Insights: from Social Psychology to Computational User Modeling

A Dissertation

Presented to

the Faculty of the School of Engineering and Applied Science

University of Virginia

In Partial Fulfillment

of the requirements for the Degree

Doctor of Philosophy (Computer Science)

by

Lin Gong

August 2019

APPROVAL SHEET

The dissertation

in submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Lin Gong

AUTHOR

The dissertation has been read and approved by the examining committee:

Hongning Wang (Advisor)

Mary Lou Soffa (Chair)

Nikolaos Sidiropoulos

Yanjun Qi

Quanquan Gu

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, Dean, School of Engineering and Applied Science

August

2019

Abstract

In the field of social psychology, a huge amount of research has been conducted to understand human behaviors by studying their physical space and belongings. Inspired from these fruitful findings, can we design corresponding computational models to characterize diverse online user behaviors by exploring users' behavior data, to understand diverse user intents? Can we further integrate different types of behavior signals driven by the same intents, to build a unified model for each user? Thanks to the advent of participatory web, which created massive amounts of user-generated data, we are able to study online user attributes and behaviors from these clues. Traditional social psychology studies commonly conduct surveys and experiments to collect user data in order to infer attributes of individuals, which are expensive and time-consuming. In contrast, we aim to understand users by building computational user models automatically, thereby to save time and efforts. And the principles of social psychology serve as good references for building such computational models.

In this dissertation, we develop new techniques to model online user behaviors based on user-generated data to better understand user preferences and intents. We get inspired from social psychology principles in user behavior modeling and these developed computational models provide alternative to explain human behaviors in the physical world. More specifically, we focus on two challenges: (1) model users' diverse ways of expressing attitudes or opinions; (2) build unified user models by integration of different modalities of user-generated data.

To tackle the challenge of capturing users' diverse opinions, we borrow the concept of social norms evolution to achieve personalized sentiment classification. By realizing the consistency existing in users' attitudes, we further perform clustered model adaptation to better calibrate such opinion coherence. To understand users from a comprehensive perspective, we utilize different modalities of user-generated data to form multiple companion learning tasks, which are further paired to accommodate the consistency existing in multi-modal user-generated data. And each individual user is modeled as a mixture over these paired instances to realize his/her behavior heterogeneity. To better characterize the correlation among different modalities of user-generated data, joint learning of different embeddings, together with explicit modeling of their relationships are performed, in order to achieve a comprehensive understanding of user intents and preferences. This dissertation borrows principles from social psychology to better design effective computational user modeling. It also provides a foundation for making user behavior modeling useful for many other applications as well as offers new directions for designing more powerful and flexible models.

Acknowledgements

The time spent at the University of Virginia is a precious, enjoyable, and remarkable journey while it is never easy. I cannot make it without support and help from a lot of amazing people.

First and foremost, I would like to express my sincere thanks to my advisor Dr.Hongning Wang, for his continuous support, supervision and guidance during the past years. I was confused about the research direction at the earlier years of the Ph.D. study and Dr.Wang encouraged me to explore different possibilities and gradually led me to find the research direction I am interested in and enthusiastic about. Dr.Wang himself acts as the role model for me in different aspects: his enthusiasm about research, his insights and vision about the area, his patience in mentoring students, his rigorous scientific attitudes and his diligence. Dr.Wang is supportive for every decision I made and it is really my fortune for have him as my advisor. I would like to thank my dissertation committee members, Prof. Mary Lou Soffa, Prof. Nikolaos Sidiropoulous, Dr. Yanjun Qi, Dr. Guanguan Gu, for all their helpful feedback and suggestions on my Ph.D. research and dissertation work. Especially, I want to thank Prof. Mary Lou Soffa for her generous support and help at every difficult moment I had. She has always been inspiring and motivating, encouraged me to overcome obstacles and move forward to pursue my goal.

Many thanks to my collaborators, Lu Lin, Benjamin Haines, Mohammad Al Boni, Bingrong Sun and Diheng Zhang, who gave a lot of insightful suggestions and inspirations in conducting research. I also want to thank all HCDM group members, who have built a great environment for research. I have benefited a lot from discussions with them.

During the time at the University of Virginia, I am grateful to have many friends who are supportive, smart, and diligent. They are Chunkun Bo, Beilun Wang, Weilin Xu, Chong Tang, In Kee Kim, Bo Yang, Ge Song, Jinghe Zhang, Xiaoqian Liu, Kevin Leach and so many friends that I cannot list them all. Thank you all for your friendship and support.

Finally and above all, I owe my deepest gratitude to my family, my husband Xian, my parents and my grandma. I would like to thank my husband, for his abiding love, support, patience, and understanding all the time. And I want to thank my parents and my grandparents for fostering my curiosity, supporting my decisions and showering me unconditional love and care. This dissertation is dedicated to them.

Dedicated to my family.
Xian, Mom, Dad, Grandparents — I love you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The advent of participatory web has transformed our way of living, such as how we communicate with friends, how we make purchases, how we look for romantic partners, and how we stay in touch with the world. Thanks to the rapid development of the Web, vast amounts of observational data such as opinionated text content and social interactions are generated, to enable the discovery of new knowledge about the human behaviors and attributes.

Online, users interact with various service systems to fulfill their idiosyncratic intents [1–4], and create massive user-generated data at the same time, which leaves "clues" that can be examined to infer the attributes and preferences of individual users. Such collection of personal data associated with a specific user serves as great resources to build up a conceptual understanding of the user, i.e., the user profile which indicates certain characteristics about an individual user. And the accurate depiction of user profile lays the foundation of adaptive changes to the system's behavior, with the process of obtaining such user profile known as user modeling. Without such accurate user modeling, service systems can hardly capture user needs and desires, thus to provide exactly what the user cares about. Therefore, computational user modeling is vital in helping automated systems to learn precise user profiles, thus to better address users' specific needs and to create compelling experience for individual users. That is, the systems can "do the 'right' thing at the 'right' time in the 'right' way" to individual. However, users' intents are diverse and not directly observable, which pose challenges for user modeling to capture such distinction among individuals explicitly.

Human behaviors have been studied in social psychology for a long time to understand how people think about, influence, and relate to others [5] in psychical world. Thus, the understanding of people by studying their physical space and belongings naturally serves as good references for performing computational user modeling of online users. For instance, the causes and consequences, the contexts that shape it, the evolutionary and developmental trajectories, the collective dynamics of diverse behaviors, all serve as principles and insights for modeling online user behaviors. And the massive user-generated data further provides resources to perform computational user modeling effectively, in order to understand user intents and preferences automatically.

In this dissertation, we accomplish the goal of ***understanding user intents*** via a computational perspective. As life becomes increasingly digit, observational data becomes increasingly computational - behaviors, opinions. Not only computational techniques are required to efficiently analyze the vast amounts of data, but also a computational learning framework is necessary for proper understanding and analyzing, with inspirations from social psychology. And the learning framework automates the ***learning process of user profiles*** to capture user preferences and intents by taking advantages of user-generated massive data and computational models, via the interactions between service systems and online users.

In this chapter, we first describe the motivation and overview of this dissertation in Section 1.1, and then introduce the definition of the general learning problem in Section 1.2, and discuss the organization of this dissertation in Section 1.3.

1.1 Motivation and Overview

User modeling builds up a conceptual representation of users, and thus, 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 [6, 7]. Due to the distinct personal characteristics and perceived environments, there exists a diverse range of human needs and desires among individual users, leading to varying decision making autonomy. Even for the same task, different users may exhibit distinct preferences and interests in order to meet their unique needs, which makes a population-level solution insufficient to address the diversity existing among users. Thus, it calls for personalized user models to capture individual users' diverse information needs accurately and effectively, so as to assist service systems to provide information the users care about or interested in, quickly and efficiently.

1.1.1 Motivating Example

For example, a student is looking for ideal jobs on professional network websites. As he/she browses a set of potential companies and jobs, the system should be smart enough to know the student's preferences regarding to location, job duty, salary, environment, growth path, training opportunity, and promotion mechanism, by retrieving all available information provided by the student, such as the text posts, the skill sets, the working experience and so on. Moreover, the connections also play an important role as they can help reveal the traits of character of the student, discover the communities the student belongs to, find jobs or companies the student interested in. Through such in-depth analysis of available information, the system can learn the particular student's job preferences, e.g. he/she emphasizes much more on the growth path and training opportunity than salary or location. Due to the uniqueness of the student's preferences, a population-level model can hardly capture the student's desires accurately and a personalized model is indeed necessary to address the issue. By knowing exactly the student's preferences, the system can recommend the appropriate jobs for the student to apply for, which helps the student find the perfect match quickly and accurately and make a good start for his/her career. Beside, the precise capture of needs for both job hunters and recruiters can quickly establish the invisible bridges between them, which is beneficial for both sides.

The example above illustrates the picture of computational user modeling studied in this dissertation. First, the intelligent system should be able to identify ***individual user's preferences*** beyond the global preferences, such as the user cares more about growth path than location; second, the system should utilize ***all available information generated by the user*** to enhance the understanding of the user's intents and preferences as they are driven by the same person and thus are consistent after all. Such in-depth understanding of user will help the service systems to accurately recognize various information needs, capture the correlations among their distinct information seeking behaviors, and optimize the quality of delivered information.

Numerous successes have proved the value of user behavior modeling in practical applications. For example, many researchers performed user behavior modeling to enable the applications in the filed of health care. For instance, Rivera-Illingworth et al. [8] presented a novel embedded agent mechanism to build normal user behavior models by recognizing their activities inside an environment recorded by unobtrusive sensors and effectors to warn signs of Alzheimer and dementia. Christakis et al. [9] analyzed a densely interconnected social network consisting of 12,067 people assessed repeatedly from

1971 to 2003, to understand the relationships between users' social behaviors and the spread of Obesity. They indeed find that network phenomena is closely relevant to the biologic and behavioral trait of obesity, and obesity appears to spread through social ties, which provides significant implications for clinical and public health interventions. Ma et al. [10] incorporated discrete prior medical knowledge of patients to their characteristics via posterior regularization, to predict patients' potential diseases more accurately. Successful user modeling also brings in tremendous business values for online information systems. For example, Zhang et al. [3] found that modeling users' review content for explainable recommendation improved CTR by more than 34.7% and conversion rate by more than 25.5% in an online e-commerce website; Liu et al. [11] reported that the frequency of website visits in Google News can be improved by more than 14.1% by modeling users' genuine topical interests on news articles and the influence of local news trends; Ai et al. [12] found that in the task of personalized product search on Amazon, learning distributed representations for users, together with queries and products, can help improve the MAP as high as 325%.

1.1.2 Challenges

Various behavior signals have been explored with different focuses in exploiting information about users' intents, such as log files [13, 14], video service [15] and network structure [16, 17]. Among all the diverse means, user-generated data is exceptionally powerful and useful as it reveals the nature of users and gives accurate insights into what really matters to them. People are actively involved in different online platforms and want their voices to be heard through the power of a line or two of text, a small, filtered image, or a low-resolution video slice. Take a closer look at the user-generated data, it is easy to find that people are talking about products or services, sharing their interests, seeking like-minded individuals. That is, they are expressing themselves - their preferences toward other items or people. Therefore, studying such data usually helps reveal the inner desires of users extensively. And we especially focus on two typical types of user-generated data: text content that indicates users' topical interests and attitudes; network structure that depicts user connectivity, as they are two most widely available and representative forms of user-generated data.

Though massive amount of data is generated among the whole population, still, a large portion of users have limited amount of data for deep exploration. Thus, most solutions on user modeling fall into the category of population-level analysis [18–20] due to the sparsity of individual user's data, which is insufficient to capture the nuance among different users. Due to the unique personality and growing environment, each individual owns his/her preferences towards the world. Even for the simple task of expressing opinions, different users may show quite different patterns, such as they may use the same word for totally different attitudes, or use different words to express the same opinions. ***Thus, simple population-level solutions would lose the resolution for precise understanding of each individual user***.

To capture individual intents via user-generated data, much efforts are devoted to explore or retrieve one specific type of user-generated data, e.g., textual content, or network structure, to understand one particular type of user behaviors. However, data usually comes with different modalities which carry different information. For example, it is very common users share a set of friends beside the textual posts or reviews, to convey their social intents. Thus, it is insufficient to characterize users' diverse information needs and desires with a single type of data. And it is important to develop a novel model which is able to jointly represent the information such that the relationships between different modalities can be captured. That is, exploring multiple modalities of user-generated data to understand user preferences from a holistic view. Essentially, different modalities of data is generated by the same user, thus, is consistent and united. ***The lack of such multi-modal modeling would lose resolution for analyzing user behaviors, leading to one-sided user understanding.***

To overcome the aforementioned limitations, more thorough and comprehensive user behavior modeling principles and approaches are needed. However, the task is never trivial because of three major concerns. ***First, user-generated data is noisy, incomplete, highly unstructured, and tied with social interactions*** [21], which imposes serious challenges in modeling such data. For example, in an environment where users are connected, e.g., social network, their generated data is potentially related, which directly breaks the popularly imposed independent and identically distributed assumptions in most learning techniques [22–24]. ***Second, different users' intents are diverse and dynamic***, which may diverge a lot among the whole population of users and may evolve over time person by person. For instance, in expressing attitudes, "expensive" may indicate negative feeling in comments like "the item is too expensive" while it may depict positive feeling in comments like "the expensive cellphone is worth the price". And accurate distinction of such nuance is of great importance. ***Third, though oftentimes scattered and sparse, such observations are neither isolated nor independent***; indeed, they reflect users' underlying intents as a whole and requires the consideration of corresponding interactions. Just as described in the proverb "Birds of a feather flock together", users of similar interests and preferences tend to be friends, indicating

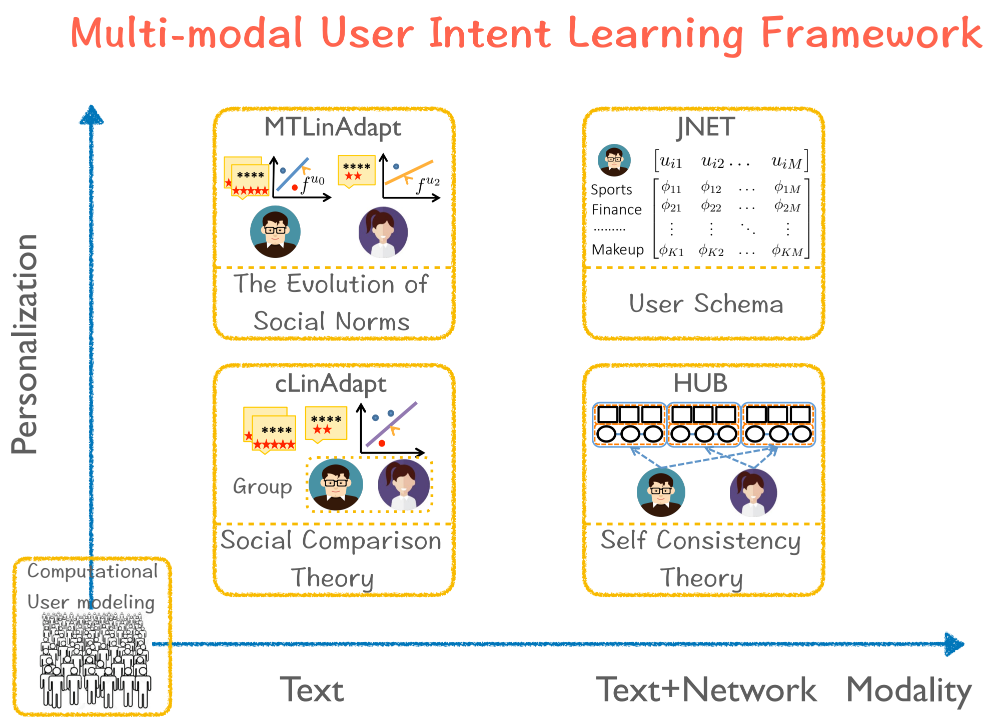

Figure 1.1: Overview of the learning framework proposed in the dissertation

the consistency existing among different types of observations.

To tackle these challenges, it is inevitable to perform effective computational user modeling to capture user intents. More importantly, we seek a comprehensive and unique solution for each individual user by modeling each user from a holistic view, i.e., explore all information available and the corresponding relationships. Thus, we propose ***a multi-modal user intent learning framework*** to address the challenges, which is illustrated in Figure 1.1. We start from mining one particular modality of user-generated data, text content, to understand how users express attitudes differently. To encode the data sparsity issue, we perform linear transformations over global parameters to alleviate the problem. By noticing the fact that similar users tend to share similar opinions, we further cluster like-minded individuals to better address the sparsity issue. On the other dimension of modality, we further incorporate user-generated network structure to capture user intents from a comprehensive perspective. Therefore, joint modeling of both modalities is conducted with both implicit and explicit capture of the correlations between them, to achieve the goal.

1.2 Problem Formation

In the section, we formally discuss the multi-modal user intent learning framework for user behavior modeling in the dissertation, with specifying the input, output and the corresponding computational algorithms.

1.2.1 Multi-modal User Behavior

Human behavior is the response of individuals to internal and external stimuli. Thus, we interpret online user behavior as a set of observations resulted from the interactions with the online service systems under one specific task T, such as writing a textual review for one particular product on crowd-sourced review forums, searching for machine learning tutorials on video sharing platforms or purchasing the product based on recommendation provided by eCommerce websites. We further quantify a particular observation $\boldsymbol{x}_d$ as a M-dimensional vector characterized by corresponding feature set $\{f_1, f_2, .., f_M\}$, which is generated by user u_i in interacting with the service system.

As one major type of behaviors studied in the dissertation, ***writing text reviews for expressing attitudes*** is a typical type of user behaviors, which creates massive amount of opinionated text data. The corresponding observation $\boldsymbol{x}_d$ can then be characterized by selected representative textual features, generating a M-dimensional vector. Text reviews usually come with numerical ratings indicating the direct attitudes of the users, which can be treated as the behavior label for the observation. Thus, we formally define y_d as the label for the observation $\boldsymbol{x}_d$, which is an observable numerical variable, either discrete or continuous. And each observation consists of two parts, the observation content and observation label, i.e., $(\boldsymbol{x}_{i,d}, y_{i,d})$.

Another extensively studied behavior in this dissertation is ***social behavior***, i.e., whether the user u_i makes friends with the other users $\{u_j\}_{j \neq i}$. The corresponding observation between a pair of users (u_i, u_j) is indeed a one-dimensional scalar with the value being the affinity between them. The social behavior of the user u_i can also be interpreted as a vector with the dimension being the number of all the users and each element is the affinity between the current user u_i and the other user u_j, i.e., $\{(u_i, u_j)\}_{j=1}^{U}$.

We should note that the behavior labels might not exist, such as the observed social connections or posted forum discussions. Also, they might not be directly available such as the comments for news articles or posts on microblogs, in which case human-annotation is needed. Usually, each user is involved with a set of observations, which can be denoted as $\{X, Y\} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_d, y_d)\}$.

Online users are usually associated with multiple behaviors resulted from different tasks, generating multiple modalities of data. Due to the distinct statistical properties of different information resources, it is important to discover the relationships among different modalities. Thus, we introduce the concept of ***multi-modal user behavior*** to act as different types of observations possessed by a particular user. Especially, we focus on ***user-generated text and network*** in the dissertation as they are two most available and representative user-generated data sources in inferring user intents.

1.2.2 Diverse User Intents

User intents depict what a user wants and usually results in intentional actions. Such intentional actions are functions to accomplish desired goals and are based on the belief that the course of actions will satisfy the desires [25]. Due to the diversity existing among users, we cannot learn global user intents as a whole. Instead, we focus on each individual user's own intents based on his/her own observations. Thus, user modeling aims at capturing each user's intents by building up a conceptual representation of the user, thus to explain observational behaviors or actions of the user. More specifically, we interpret the intents as the preferences towards the target attributes. Assume user profile is a M-dimension vector $\boldsymbol{u} = [\boldsymbol{u}_1, \boldsymbol{u}_2, .., \boldsymbol{u}_M]$, the target attributes are a L-dimension vector $\boldsymbol{p} = [\boldsymbol{p}_1, \boldsymbol{p}_2, .., \boldsymbol{p}_L]$ with each element indicting one basic attribute, user intents are then quantified as a L-dimensional vector $\boldsymbol{h} = [\boldsymbol{h}_1, \boldsymbol{h}_2, .., \boldsymbol{h}_L]$ with each dimension representing the emphasis on the corresponding attribute. In order to align user profile with user intents, a proper mapping function is needed to indicate the proximity between the user and the target attributes, i.e., $\boldsymbol{h} = f(\boldsymbol{u}, \boldsymbol{p})$. For instance, in the task of learning social intents, the basic attribute can be considered as every other user, the user intents are then quantified as the affinity between the current user and every other user, via defining mapping function as one affinity calculation method between a pair of user vectors. ***Formally, we capture user intents by learning user emphasis on target attributes via proper mapping of user profile, based on the user's observations in specific tasks.***

Moreover, user intents vary among different tasks, giving different practical meanings of user profile and corresponding mapping function. For instance, in the task of opinion mining, user intents can be

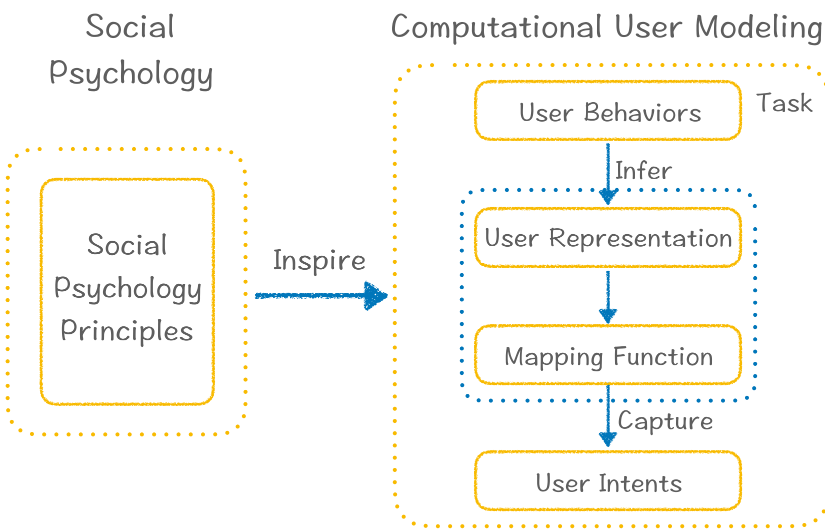

Figure 1.2: Relations of different components in the dissertation

interpreted as preferences over sentiment features, thus the learned feature weights directly act as the user representation. While in the task of joint modeling of textual interests and social intents, the learned user vector acts as a more general representation conveying consistency inside each user's mental states, and the preferences regarding to different attributes are realized by different mapping functions with the same input user profile.

The problem studied in the dissertation is to infer the user intents by learning a unified user representation, together with the corresponding mapping functions, which best explain the diverse sets of observations, i.e., $g : X \to H \sim f(U, P)$, where X could take different forms of user-generated data and mapping function f could also adjust based on observations. Moreover, principles from social psychology provide insights in designing such computational framework such as imposing assumptions and quantifying concepts or states. We explain the relations among different components in Figure 1.2 to better illustrate the problem defined.

1.2.3 User Intent Learning

By defining the input and output of the problem studied in the dissertation, we can formalize the computational problem in a principled way: Given a user's multi-modal observations, learn user intents characterized by mapping from user representation $\boldsymbol{u}$ to attributes $\boldsymbol{r}$, that can best explain the user's different sets of observations, as show in Eq 1.2.1.

$$\operatorname*{argmax}_{H} p(X|H)p(H|f(U,R)) \tag{1.2.1}$$

where X is the observations, H and U are user intents matrix and user profile matrix among all the users.

Note, the mapping function could take different forms and could be more than one. If it is a direct mapping from user representation to the target attributes, the mapping function can be defined as simple as a constant. Assume the user representation shares the same space with the attribute unit, then the representation itself already reveals the preferences, thus can be directly understood as the user intents. For instance, in the task of opinion mining, we embed user representation to the same space of sentiment features, thus the learned feature weights reveal the emphasis on sentiment units, thus naturally serve as the user representation. However, if there are multiple sets of attributes in different space, then multiple mapping functions are needed to encode the distinct emphasis towards each individual attribute space.

1.3 Dissertation Organization

In the dissertation, we focus on modeling user behaviors by utilizing user-generated data, to capture diverse user intents. More specifically, in the first part of ***Modeling Opinionated Text for Personalized Sentiment Analysis***, we aim at learning user profile characterized by users' ways of expressing opinions via utilizing diverse textual information indicating users' true feelings and interests; in the second part of ***Incorporating Network for Holistic User Behavior Modeling***, we further incorporate users' social connections to perform a multi-modal learning to encode the correlations among multiple modalities, thus, to gain a comprehensive understanding of user preferences as a whole. An overview of the dissertation is described in Figure 1.1 to better illustrate our goal.

•Chapter 3: Modeling Social Norms Evolution for Personalized Sentiment Classification
Expressing attitudes is a typical type of user behaviors, indicating users' preferences toward sentiment words. And user-generated textual information provides great resources to examine such behavior to understand user intents, which is also known as sentiment analysis or opinion mining. Sentiment is personal as the same sentiment can be expressed in various ways and the same expression might

carry distinct polarities across different individuals [26]. The sparsity of individual user's data limits the exploration of his/her attitudes. Thus, current mainstream solutions of sentiment analysis usually focus on population-level models with two typical types of studies [27, 28]. The first is classifying input text units (such as documents, sentences and phrases) into predefined categories, e.g., positive v.s., negative [18, 29] and multiple classes [20]. The second is identifying topical aspects and corresponding opinions, e.g., developing topic models to predict fine-grained aspect ratings [6, 30]. All these works emphasize population-level analysis which apply a global model on all users, due to the limited amount of observations of individual users. Therefore, such solution fails to recognize the heterogeneity in which different users express their diverse opinions. Instead, we want to capture individual users' diverse ways of expressing attitudes by overcoming the sparsity issue.

As sentiment analysis is extensively studied in social science, we get motivated by the finding that people's opinions are diverse and variable while together they are shaped by evolving social norms. In Chapter 3, we perform ***personalized sentiment classification via shared model adaptation*** over time. In our proposed solution, a global sentiment model is constantly updated to capture the homogeneity in which users express opinions, while personalized models are simultaneously adapted from the global model to recognize the heterogeneity of opinions from individuals. Global model sharing alleviates data sparsity issue, and individualized model adaptation enables efficient online model learning to realize the intent learning in expressing attitudes.

•Chapter 4: Clustered Model Adaptation for Personalized Sentiment Analysis

With the aforementioned personalized sentiment analysis, little performance improvement can be achieved for users with limited amount of observations while they form a major portion of the user population. Thus, we take a new perspective to build personalized sentiment models by exploiting social psychology theories about humans' dispositional tendencies. Suggested by the *Theory of Social Comparison* [31], 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. Therefore, in Chapter 4, we propose to ***capture such clustering property of different users' opinions*** by postulating a non-parametric Dirichlet Process (DP) prior [32] 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). According to the *Cognitive Consistency Theory* [33], 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 [34]. 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 norms, because it is estimated based on observations from all users. It thus captures homogenous sentimental regularities across the 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.

•Chapter 5: A Holistic User Behavior Modeling via Multi-task Learning

The availability of online social network provides an opportunity to learn users from extrinsic regulations, i.e., the implicit social influence, in addition to self regulations. Nowadays, a lot of efforts are devoted to network structure analysis. For instance, the proximity between a pair of users has been studied to understand social influence and information diffusion [22]; and network structure has been analyzed to examine users' social grouping and belongings [35–37]. These works restrict the analysis within network structure and fail to realize the dependency among different types of user-generated data. That is, the consistency between extrinsic regulations and self regulations is ignored, which is essentially governed by the same user. As the influence is not visible and measurable, it is difficult to quantify the influence itself and the impacts of influence to user modeling.

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 supported by the *Self Consistency Theory* [38] in social psychology study, as it asserts that consistency of ideas and representation of the self are integral in humans.

In Chapter 5, we focus on user modeling in multiple modalities of user-generated data, where users write text reviews 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., a multi-task learning approach [39–41]. The two tasks are paired to encode the consistency existing in user intents. Instead of assigning one paired task for each user, a specific set of such paired instances are assumed to accommodate the homogeneity among users' behaviors. Individual users are then modeled as a mixture over the unique instances of paired learning tasks to realize his/her own behavior heterogeneity.

•Chapter 6: User Representation Learning with Joint Network Embedding and Topic Embedding

In order to better understand user intents, ***explicitly modeling the structural dependency among different modalities*** of user-generated data is of vital importance. This statement is also supported by existing qualitative studies in social psychology, i.e., the concept the *User Schema*. User Schema defines a generalized representation for understanding the knowledge of a person, which organizes: *1) categories of information*; and *2) the relationships among them.* Thus, people's online schema enabled by their large amounts of online behavior data, naturally fits our goal of learning unified user representation. And it further motivates us to learn such representation in a shared latent space, and the concept of distributed representation learning [42], i.e., user embedding, provides the concrete solution. The user representation is learned in a low-dimensional continuous space, where the structural dependency among different modalities of user-generated data can be realized by the proximity between the users and their generated data.

In Chapter 6, we utilize text content and social interactions for user representation learning. We develop a probabilistic generative model to integrate user representation learning with content modeling and social network modeling. On the one hand, we embed topics into the same latent space with users to model user-generated text content. A user's affinity to a topic is characterized by his/her proximity to the topic in this learned space. A user's text document is then generated with respect to the projected topic vectors on his/her user embedding vector. On the other hand, the affinity between users reflected in their social interactions is directly modeled by the proximity between users' embedding vectors. The observed network edges are sampled from this underlying distribution of user affinity. The user representation is obtained via posterior inference over a set of training data, which can be efficiently performed via a variational Bayesian procedure.

To sum up, we explored users' diverse intents reflected in their corresponding behaviors by building computational user models. Instead of population-level user modeling, we performed individual-level user understanding via personalized learning. We borrowed principles and concepts in social psychology to facilitate the design of the computational models, which bridges the gap between the two communities. More importantly, multiple modalities generated by the users are integrated in different ways to achieve a comprehensive understanding of the user intents.

Chapter 2

Background

2.1 Sentiment Analysis

Text mining, also known as text analytics, aims at generating structured data out of free text content to extract machine-readable facts from them. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling

Sentiment analysis, also called opinion mining, analyzes peoples opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [27]. Although linguistics and natural language processing (NLP) have been studies for a long time, little attention had been paid on peoples opinions and sentiments before the year 2000. Sentiment analysis becomes possible and prevalent later on due to two major reasons: 1) the constantly growing popularity and availability of opinion-rich textual content such as personal blogs and online reviews enable the study of people's opinions and attitudes; 2) the wide range of applications in diverse domains such as political science, economics, and social sciences, offer many challenging problems and motivate the research.

The primary goal of sentiment analysis includes data analysis on the body of the text for understanding the opinion expressed by it and other key factors comprising modality and mood. Text-based sentiment classification forms the foundation of sentiment analysis [27, 28]. A great deal of efforts have been devoted in the exploration of opinion-rich textual content to understand users' decision making process [20, 27, 28].

There are two typical types of studies in sentiment classification. The first is **classifying input text units into predefined categories**, such as classifying documents, sentences and phrases into positive/negative [18, 29] or multiple classes [20]. Both lexicon-based [43–45] and learning-based [20, 28] solutions have been explored. For instance, [43] mined the features of the product on which the customers have expressed their opinions. Then it concluded whether the opinions for a particular feature are positive or negative by inferring the corresponding sentiment of the sentences containing the feature. Different from the perspective of lexicon-based methods, Pang et al. [18] examined several supervised machine learning methods for sentiment classification for the first time, on the set of movie reviews and concluded that machine learning techniques outperform the method that is based on human-tagged features.

The second is **utilizing the user-generated text data to understand users' emphasis on specific entities or aspects** [2, 6, 30, 46]. Statistical topic models [47, 48] serve as a building block for statistical modeling of text data. Typical solutions model individual users as a bag of topics [49], which govern the generation of associated text documents. Wang and Blei [50] combine topic modeling with collaborative filtering to estimate topical user representations with additional observations from user-item ratings. Wang et al. [2] analyzed users' opinions about entities in an online review at the level of topical aspects to discover each individual reviewers latent opinion on each aspect, together with their emphasis on those latent aspects.

However, those works either emphasize population-level or document-level analysis while users' feelings and attitudes are personal. Thus, population-level analysis which applies a global model on all users, cannot recognize the heterogeneity in users' different ways of expressing opinions. Though document-level analysis treats each document individually, it may ignore the consistency existing in each individual user's documents. To accommodate the heterogeneity among users and the homogeneity inside each individual user, user-level sentiment analysis is necessary to achieve the goal.

Sparse observations of individuals opinionated data [51] prevent straightforward solutions from building personalized sentiment classification models, such as estimating supervised classifiers on a per-user basis. Semi-supervised methods are developed to address the data sparsity issue. For example, leveraging auxiliary information from user-user and user-document relations in transductive learning [23, 52]. However, only one global model is estimated there, and the details of how individual users express diverse opinions cannot be captured. Our works overcome the limitations by building personalized sentiment classification models through shared model adaptation. Instead of building personal sentiment models from scratch, a predefined global sentiment model serves as the basis model for diverse users to adapt from, so as to alleviate the data sparsity issue. In order to minimize the efforts of realizing the personalization, linear transformations are utilized to realize the adaptation from global model to personal models.

Moreover, the development in modeling user-generated text data directly **enables personalized recommendation and retrieval**. Zhang et al. [3] combined phrase-level sentiment analysis with matrix factorization for explainable recommendation. Ghose et al. [53] illustrated how user-generated content can be mined and incorporated into a demand estimation model so as to generate a new ranking system in product search engines.

2.2 Social Network Analysis

Social Network Analysis has gained increasing importance in recent years. It is the process of investigating social structures through the use of networks and graph theory [54], which has been used extensively in a wide range of applications and disciplines, including link prediction, community detection, network propagation modeling and so on [55].

The task of **link prediction** predicts missing links in current networks and new or dissolution links in future networks, which plays an important role in mining and analyzing the evolution of social networks. Among the various methods realizing link prediction, they can roughly fall into two categories. The first category of methods utilize the neighbor-based metrics to infer the missing links [22, 56], where different similarity functions are used to achieve so. Different similarity measures are extensively evaluated by Liben-Nowell and Kleinberg [22] who found that the Adamic-Adar measure of node similarity performed best. Cha et al. [57] interpret the directed links as the flow of information and hence indicate a users influence on others. Correspondingly, they presented an in-depth comparison of three measures of influence: indegree, retweets, and mentions and discovered many interesting facts about the dynamics of user influence across topics and time. Huo et al. [58]

calculated the linking probability between a pair of users by further considering their activities. The second category of methods employ path-based metrics for link prediction, in which random walk is designed to traverse the paths between two nodes to calculate the proximity. Tong et al. [59] studied the role of directionality in measuring proximity on graphs. They defined a direction-aware proximity measure based on the random walk notion of escape probability, which naturally weights and quantifies the multiple relationships reflected through the many paths connecting node pairs. Backstrom et al. [60] utilized node and edge attribute data to guide the random walks towards the desired target nodes.

Another important task in studying networks is **identifying network communities**. Fundamentally, communities allow us to discover groups of interacting objects (i.e., nodes) and the relations between them. Thus, identifying network communities can be viewed as a problem of clustering a set of nodes into communities, where a node may belong to multiple communities. Typically, two sources of data can be used to perform the clustering task, the node attributes and network structure. Therefore, some clustering methods focus on identifying nodes sharing similar attributes [36, 61]. Wang et al. [61] assumes that each node belongs to a cluster and the relationships between nodes are governed by the corresponding pair of clusters. With posterior inference, one identifies a set of latent roles which govern the nodes relationships with each other. Airoldi et al. [35] further extended the single membership of each node to mixed membership, which provides more flexibility. There are also some works aiming to find communities based on the network structure, e.g., to find groups of nodes that are densely connected [62, 63]. Though most works utilize one source to detect communities,Yang et al. [64] developed an accurate and scalable algorithm for detecting overlapping communities in networks communities with both Edge Structure and Node Attributes considered. Due to the modeling of both sources of data, i.e., the interaction between the network structure and the node attributes, it leads to more accurate community detection as well as improved robustness in the presence of noise in the network structure.

Considerable efforts are made in utilizing network structure for learning more concise and effective user representations, to facilitate the advanced analytic tasks. **Network embedding** techniques [17, 65], which assigns nodes in a network to low-dimensional representations and effectively preserves the network structure, naturally fit the needs of concise and effective user representation. Recently, a significant amount of progresses have been made toward this emerging network analysis paradigm. Inspired from word embedding techniques [66], random walk models are exploited to generate random paths over a network to learn dense, continuous and low-dimensional representations of users [16, 17, 67]. Perozzi et al. [16] claims that the vertex frequency in random walks on scale free graphs also follows a power law. Thus the truncated random walks are treated as sentences for Skip-gram to learn the corresponding user embeddings. Grover et al. [67] proposed a framework, node2vec, to learn low-dimensional representations of nodes in a graph by optimizing a neighborhood preserving objective. With a flexible objective, the algorithm accommodates for various definitions of network neighborhoods by simulating biased random walks. In the work [17], a novel network embedding method LINE is developed to analyze arbitrary types of information networks: undirected, directed, and/or weighted. It optimizes a carefully designed objective function that preserves both the local and global network structures. Matrix factorization technique is also commonly used to learn user embeddings [68, 69], as learning a low-rank space for an adjacency matrix representing the topology of a network naturally fits the need of learning low-rank user/node embeddings. For instance, Wang et al. [69] propose a modularized non-negative matrix factorization model to incorporate the community structure into network embedding. Tang and Liu [70] factorize an input network's modularity matrix and use discriminative training to extract representative dimensions for learning user representation. Deep neural network based methods [71–73] are also proved to be effective in learning node representations as they can perform non-linear mapping between original space and embedding space.

2.3 Joint Modeling of Text and Network

Though much efforts are devoted to the exploration of either text data or network structure, little attention is paid to the joint modeling of the two modalities. Since they are generated by the same person, there exists consistency between them. Due to its statistical property, they may also complement each other. There are indeed some works combine text content with network structure to improve the fidelity of learned user models. Speriosu et al. [74] proposed to propagate labels from a supervised classifier over the Twitter follower graph to improve sentiment classification. Tan et al. [23] believe that connected users are more likely to hold similar opinions; therefore, relationship information can be incorporated to complement extraction about a users viewpoints from their utterances. Hu et al. [75] developed a novel sociological approach to handle networked texts in microblogging. In particular, they extracted sentiment relations between textual documents based on social theories, and model the relations using graph Laplacian, which is employed as a regularization to a sparse formulation. Cheng et al. [76] leveraged signed social network to infer the sentiment of text documents in an unsupervised manner. Tang et al. [77] 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. Pozzi et al. [78] utilized the approval relations to estimate user polarities about a given topic in a semi-supervised framework. Joint modeling of text and network is also enabled in the framework of matrix factorization. By proving that DeepWalk is actually equivalent to matrix factorization (MF), Yang et al. [79] further incorporated text features of vertices into network representation learning.

However, all the aforementioned works only treat the network as side information for text data modeling, and they do not model the interactions between different modalities of data. Our works unify the modeling of textual data and network structure to capture the relationship between the two modalities, thus to enable a comprehensive understanding of user intents.

2.4 Multi-task Learning

Learning multiple related tasks simultaneously has been empirically [80–85] as well as theoretically [86–88] shown to often significantly improve performance relative to learning each task independently. Multi-task learning exploits the commonalities and difference across tasks to facilitate the information sharing. 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 [39, 89]. This assumption has been empirically proved to be effective for capturing preferences of individual users [82]. Evgeniou et al. [39] first generalized the regularizationbased methods from singletask to multitask learning, which are natural extensions of existing kernel based learning methods for single task learning. Task relatedness has also been imposed via constructing a common underlying representation across different tasks [87, 88, 90, 91]. For instance, in modeling users preferences/choices, it may also be the case that people make decisions (e.g. purchase of cellphones, visit of restaurants, etc.) using a common set of features describing these products or service business. [91] presented a method for learning a low-dimensional representation which is shared across a set of multiple related tasks.

Building personalized user models can be considered as a multi-task learning problem as users are correlated to certain extends. By exploiting the relatedness among users/tasks, they can reinforce each other mutually. Similar modeling approaches have been explored in user modeling before. Fei et al. [92] 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.

2.5 User Behavior Study in Social Psychology

Individual behavior has been studied in social psychology for a long time. Kurt Lewin, who is recognized as the "founder of social psychology", points out that behavior is affected both by the individual's personal characteristics and by the social environment as he or she perceives it [93]. As an important component of social psychology, behavior is extensively studied from both the individual perspective and social perspective. That is, social psychologists study how people view themselves and each other, how they interpret people's behaviors, how their attitudes form and change, how people act in groups and how groups affect each other. In order to conduct such studies, data collection such as surveys or experiments are an important part in the process. Researchers define what they need, design questionnaires or experimental treatments, and collect the data based on the results of administering them [5]. The results provide a great degree of control over the measurement but is expensive, time consuming and may even be dangerous sometimes.

With the data explosion of recent years, more and more data is generated in various aspects of life for studying user behaviors. As we can understand people by studying their physical space and belongings, we are also able to investigate users by studying their online connections, postings and actions [5], enabled by the advent of participatory web. And the corresponding social psychological principles in psychical space naturally serve as great resources to help understand user behaviors in virtual space. In turn, the effective user modeling provides alternative or replacement for traditional research techniques in social psychology.

sIndeed, much efforts are devoted to bridging the gap between social psychology and computational user behavior modeling. One line of research focus on verifying the correlation between real-world user behaviors and online user modeling. For instance, O'Connor et al. [19] analyzed sentiment polarity of a huge number of tweets and found a correlation of 80% with results from public opinion polls. Bollen et al [94] used Twitter data to predict trends in the stock market. They showed that one can predict general stock market trends from the overall mood expressed in a large number of tweets. Another line of research gets inspired from knowledge of social psychology in building computational models of online users. Tan et al [23] and Hu et al [75] adopted the principle of homophily, i.e., "birds of a feather flock together", to utilize social connections to complement users' attitudes.

Part I

Modeling Opinionated Text for Personalized Sentiment Analysis

Chapter 3

Modeling Social Norms Evolution for Personalized Sentiment Classification

In this Chapter, we study the problem of understanding users' preferences in expressing attitudes, which is personalized sentiment classification. We get inspired from the evolution of social norms and perform personalized sentiment classification via ***shared model adaptation*** over time. In the proposed solution, a global sentiment model is constantly updated to capture the homogeneity in which users express opinions, while personalized models are simultaneously adapted from the global model to recognize the heterogeneity of opinions from individuals. Global model sharing alleviates data sparsity issue, and individualized model adaptation enables efficient online model learning.

3.1 Introduction

Sentiment is personal; the same sentiment can be expressed in various ways and the same expression might carry distinct polarities across different individuals [26]. Current mainstream solutions of sentiment analysis overlook this fact by focusing on population-level models [27, 28]. But the idiosyncratic and variable ways in which individuals communicate their opinions make a global sentiment classifier incompetent and consequently lead to suboptimal opinion mining results. For instance, a shared statistical classifier can hardly recognize that in restaurant reviews, the word "expensive" may indicate some users' satisfaction with a restaurant's quality, although it is generally associated with negative attitudes. Hence, a personalized sentiment classification solution is required to achieve fine-grained understanding of individuals' distinctive and dynamic opinions and benefit downstream opinion mining applications.

Sparse observations of individuals' opinionated data [51] prevent straightforward solutions from building personalized sentiment classification models, such as estimating supervised classifiers on a per-user basis. Semi-supervised methods are developed to address the data sparsity issue. For example, leveraging auxiliary information from user-user and user-document relations in transductive learning [23, 75]. However, only one global model is estimated there, and the details of how individual users express diverse opinions cannot be captured. More importantly, existing solutions build static sentiment models on historic data; but the means in which a user expresses his/her opinion is changing over time. To capture temporal dynamics in a user's opinions with existing solutions, repeated model reconstruction is unavoidable, albeit it is prohibitively expensive. As a result, personalized sentiment analysis requires effective exploitation of users' own opinionated data and efficient execution of model updates across all users.

To address these challenges, we propose to build personalized sentiment classification models via ***shared model adaptation***. Our solution roots in the social psychology theories about humans' dispositional tendencies [95]. Humans' behaviors are shaped by social norms, a set of socially shared "feelings" and "display rules" about how one should feel and express opinions [34, 96]. In the context of content-based sentiment classification, we interpret social norms as global model sharing and adaptation across users. Formally, we assume a global sentiment model serves as the basis to capture self-enforcing sentimental regularities across users, and each individual user tailors the shared model to realize his/her personal preference. In addition, social norms also evolve over time [97], which leads to shifts in individuals' behaviors. This can again be interpreted as model adaptation: a new global model is adapted from an existing one to reflect the newly adopted sentimental norms. The temporal changes in individuals' opinions can be efficiently captured via online model adaptation at the levels of both global and personalized models.

Our proposed solution can also be understood from the perspective of multi-task learning [39, 89].

Intuitively, personalized model adaptations can be considered as a set of related tasks in individual users, which contribute to a shared global model adaptation. In particular, we assume the distinct ways in which users express their opinions can be characterized by a linear classifier's parameters, i.e., the weights of textual features. Personalized models are thus achieved via a series of linear transformations over a globally shared classifier's parameters [1], e.g., shifting and scaling the weight vector. This globally shared classifier itself is obtained via another set of linear transformations over a given base classifier, which can be estimated from an isolated collection beforehand and serves as a prior for shared sentiment classification. The shared global model adaptation makes personalized model estimation no longer independent, such that regularity is formed across individualized learning tasks.

We empirically evaluated the proposed solution on two large collections of reviews, i.e., Amazon and Yelp reviews. Extensive experiment results confirm its effectiveness: the proposed method outperformed user-independent classification methods, several state-of-the-art model adaption methods, and multi-task learning algorithms.

3.2 Related Work

The idea of model adaptation has been extensively explored in the context of transfer learning [98], which focuses on applying knowledge gained while solving one problem to different but related problems. In opinion mining community, transfer learning is mostly exploited for domain adaptation, e.g., adapting sentiment classifiers trained on book reviews to DVD reviews [99, 100]. Personalized model adaptation has also been studied in literature. The idea of linear transformation based model adaptation is introduced in [1] for personalized web search. Al Boni et al. applied a similar idea to achieve personalized sentiment classification [101]. [102] developed an online learning algorithm to continue training personalized classifiers based on a given global model. However, all of these aforementioned solutions perform model adaptation from a fixed global model, such that the learning of personalized models is independent from each other. Data sparsity again is the major bottleneck for such solutions. Our solution associates individual model adaptation via a shared global model adaptation, which leverages observations across users and thus reduces preference learning complexity.

3.3 Methodology

3.3.1 The Evolution of Social Norms

Social norms create pressures to establish socialization of affective experience and expression [103]. Within the limit set by social norms and internal stimuli, individuals construct their sentiment, which is not automatic, physiological consequences but complex consequences of learning, interpretation, and social influence. This motivates us to build a global sentiment classification model to capture the shared basis on which users express their opinions. For example, the phrase "a waste of money" generally represents negative opinions across all users; and it is very unlikely that anybody would use it in a positive sense. On the other hand, members of some segments of a social structure tend to feel certain emotions more often or more intensely than members of other segments [104]. Personalized model adaptation from the shared global model becomes necessary to capture the variability in affective expressions across users. For example, the word "expensive" may indicate some users' satisfaction with their received service.

Studies in social psychology also suggest that social norms shift and spread through infectious transfer mediated by webs of contact and influence over time [97, 105]. Members inside a social structure influence the other members; confirmation of shifted beliefs leads to the development and evolution of social norms, which in turn regulate the shared social behaviors as a whole over time. The evolving nature of social norms urges us to take a dynamic view of the shared global sentiment model: instead of treating it as fixed, we further assume this model is also adapted from a predefined one, which serves as prior for sentiment classification. All individual users are coupled and contribute to this shared global model adaptation. This two-level model adaptation assumption leads us to the proposed multi-task learning solution, which will be carefully discussed in the next section.

3.3.2 Shared Linear Model Adaptation

In this paper, we focus on linear models for personalized sentiment classification due to their empirically superior performance in text-based sentiment analysis [18, 20]. We assume the diverse ways in which users express their opinions can be characterized by different settings of a linear model's parameters, i.e., the weights of textual features.

Formally, we denote a given set of opinionated text documents from user u as $D^u = \{(\boldsymbol{x}_d^u, y_d^u)\}_{d=1}^{|D^u|}$, where each document $\boldsymbol{x}_d^u$ is represented by a V-dimensional vector of textual features and y_d^u is the corresponding sentiment label. The task of personalized sentiment classification is to estimate a personalized model $y = f^u(\boldsymbol{x})$ for user u, such that $f^u(\boldsymbol{x})$ best captures u's opinions in his/her generated text content. Instead of assuming $f^u(\boldsymbol{x})$ is solely estimated from user u's own opinionated data, which is prone to overfitting, we assume it is derived from a globally shared sentiment model $f^s(\boldsymbol{x})$ via model adaptation [1, 101], i.e., shifting and scaling $f^s(\boldsymbol{x})$'s parameters for each individual user. To simplify the following discussions, we will focus on binary classification, i.e., $y_d \in \{0, 1\}$, and use the logistic regression as our reference model. But the developed techniques are general and can be easily extended to multi-class classification and generalized linear models.

We only consider scaling and shifting operations, given rotation requires to estimate much more free parameters (i.e., $O(V^2)$ v.s., $O(V)$) but contributes less in final classification performance [101]. We further assume the adaptations can be performed in a group-wise manner [1]: features in the same group will be updated synchronously by enforcing the same shifting and scaling operations. This enables the observations from seen features to be propagated to unseen features in the same group during adaptation. Various feature grouping methods have been explored in [1].

Specifically, we define $g(i) \to j$ as a feature grouping method, which maps feature i in $\{1, 2, \ldots, V\}$ to feature group j in $\{1, 2, \ldots, K\}$. A personalized model adaptation matrix can then be represented as a $2K$-dimensional vector $A^u = (a_1^u, a_2^u, \ldots, a_K^u, b_1^u, b_2^u, \ldots, b_K^u)$, where a_k^u and b_k^u represent the scaling and shifting operations in feature group k for user u accordingly. Plugging this group-wise model adaptation into the logistic function, we can get a personalized logistic regression model $P^u(y_d = 1|\boldsymbol{x}_d)$ for user u as follows,

$$P^u(y_d = 1|\boldsymbol{x}_d) = \frac{1}{1 + e^{-\sum_{k=1}^{K} \sum_{g(i)=k} (a_k^u w_i^s + b_k^u)\boldsymbol{x}_i}} \tag{3.3.1}$$

where $\boldsymbol{w}^s$ is the feature weight vector in the global model $f^s(\boldsymbol{x})$. As a result, personalized model adaptation boils down to identifying the optimal model transformation operation A^u for each user based on $\boldsymbol{w}^s$ and D^u.

In [1, 101], $f^s(\boldsymbol{x})$ is assumed to be given and fixed. It leads to isolated estimation of personalized models. Based on the social norms evolution theory, $f^s(\boldsymbol{x})$ should also be dynamic and ever-changing to reflect shifted social norms. Hence, we impose another layer of model adaptation on top of the shared global sentiment model $f^s(\boldsymbol{x})$, by assuming itself is also adapted from a predefined base sentiment model. Denote this base classifier as $f^0(\boldsymbol{x})$, which is parameterized by a feature weight vector $\boldsymbol{w}^0$ and serves as a prior for sentiment classification. Then $\boldsymbol{w}^s$ can be derived via the same aforementioned model adaptation procedure: $\boldsymbol{w}^s = A^s\widetilde{\boldsymbol{w}}^0$, where $\widetilde{\boldsymbol{w}}^0$ is an augmented vector of $\boldsymbol{w}^0$, i.e., $\widetilde{\boldsymbol{w}}^0 = (\boldsymbol{w}^0, 1)$, to facilitate shifting operations, and A^s is the adaptation matrix for the shared global model. We should note A^s can take a different configuration (i.e., feature groupings) from individual users' adaptation matrices.

Putting these two levels of model adaptation together, a personalized sentiment classifier is achieved via,

$$\boldsymbol{w}^u = A^u A^s \widetilde{\boldsymbol{w}}^0 \tag{3.3.2}$$

which can then be plugged into Eq (5.3.1) for personalized sentiment classification.

We name this resulting algorithm as Mutli-Task Linear Model Adaptation, or ***MT-LinAdapt*** in short. The benefits of shared model adaptation defined in Eq (3.3.2) are three folds. First, the homogeneity in which users express their diverse opinions are captured in the jointly estimated sentiment model $f^s(\boldsymbol{x})$ across users. Second, the learnt individual models are coupled together to reduce preference learning complexity, i.e., they collaboratively serve to reduce the models' overall prediction error. Third, non-linearity is achieved via the two-level model adaptation, which introduces more flexibility in capturing heterogeneity in different users' opinions. In-depth discussions of those unique benefits will be provided when we introduce the detailed model estimation methods.

3.3.3 Joint Model Estimation

The ideal personalized model adaptation should be able to adjust the individualized classifier $f^u(\boldsymbol{x})$ to minimize misclassification rate on each user's historical data in D^u. In the meanwhile, the shared sentiment model $f^s(\boldsymbol{x})$ should serve as the basis for each individual user to reduce the prediction error, i.e., capture the homogeneity. These two related objectives can be unified under a joint optimization problem.

In logistic regression, the optimal adaptation matrix A^u for an individual user u, together with A^s can be retrieved by a maximum likelihood estimator (i.e., minimizing logistic loss on a user's own opinionated data). The log-likelihood function in each individual user is defined as,

$$L(A^u, A^s) = \sum_{d=1}^{|D^u|} \Big[y_d \log P^u(y_d = 1|\boldsymbol{x}_d) + (1 - y_d) \log P^u(y_d = 0|\boldsymbol{x}_d) \Big] \tag{3.3.3}$$

To avoid overfitting, we penalize the transformations which increase the discrepancy between the adapted model and its source model (i.e., between $\boldsymbol{w}^u$ and $\boldsymbol{w}^s$, and between $\boldsymbol{w}^s$ and $\boldsymbol{w}^0$) via a L2 regularization term,

$$R(A) = \frac{\eta_1}{2} ||\boldsymbol{a} - 1||_2 + \frac{\eta_2}{2} ||\boldsymbol{b}||_2 \tag{3.3.4}$$

and it enforces scaling to be close to one and shifting to be close to zero.

By defining a new model adaptation matrix $\mathring{A} = \{A^{u_1}, A^{u_2}, \dots, A^{u_N}, A^s\}$ to include all unknown model adaptation parameters for individual users and shared global model, we can formalize the joint optimization problem in MT-LinAdapt as,

$$\max L(\mathring{A}) = \sum_{i=1}^{N} \Big[L(A^{u_i}) - R(A^{u_i}) \Big] - R(A^s) \tag{3.3.5}$$

which can be efficiently solved by a gradient-based optimizer, such as quasi-Newton method [106].

Direct optimization over $\mathring{A}$ requires synchronization among all the users. But in practice, users will generate their opinionated data with different paces, such that we have to postpone model adaptation until all the users have at least one observation to update their own adaptation matrix. This delayed model update is at high risk of missing track of active users' recent opinion changes, but timely prediction of users' sentiment is always preferred. To monitor users' sentiment in realtime, we can also estimate MT-LinAdapt in an asynchronized manner: whenever there is a new observation available, we update the corresponding user's personalized model together with the shared global model immediately. i.e., online optimization of MT-LinAdapt.

This asychronized estimation of MT-LinAdapt reveals the insight of our two-level model adaptation solution: the immediate observations in user u will not only be used to update his/her own adaptation parameters in A^u, but also be utilized to update the shared global model, thus to influence the other users, who do not have adaptation data yet. Two types of competing force drive the adaptation among all the users: $\boldsymbol{w}_s = A^s\widetilde{\boldsymbol{w}}_0$ requires timely update of global model across users; and $\boldsymbol{w}_u = A^u\boldsymbol{w}_s$ enforces the individual user to conform to the newly updated global model. This effect can be better understood with the actual gradients used in this asychronized update. We illustrate the decomposed gradients for scaling operation in A^u and A^s from the log-likelihood part in Eq (5.3.13) on a specific adaptation instance $(\boldsymbol{x}_d^u, y_d^u)$:

$$\frac{\partial L(A^u,A^s)}{\partial a_k^u}=\Delta_d^u\sum_{g^u(i)=k}\Big(a^s_{g^s(i)}w_i^0+b^s_{g^s(i)}\Big)\boldsymbol{x}^u_{di} \tag{3.3.6}$$

$$\frac{\partial L(A^u,A^s)}{\partial a_l^s}=\Delta_d^u\sum_{g^s(i)=l}a^u_{g^u(i)}\boldsymbol{w}_i^0\boldsymbol{x}^u_{di} \tag{3.3.7}$$

where $\Delta_d^u = y_d^u - P^u(y_d^u = 1|\boldsymbol{x}_d^u)$, and $g^u(\cdot)$ and $g^s(\cdot)$ are feature grouping functions in individual user u and shared global model $f^s(\boldsymbol{x})$.

As stated in Eq (3.3.6) and (3.3.7), the update of scaling operation in the shared global model and individual users depends on each other; the gradient with respect to global model adaptation will be accumulated among all the users. As a result, all users are coupled together via the global model adaptation in MT-LinAdapt, such that model update is propagated through users to alleviate data sparsity issue in each single user. This achieves the effect of multi-task learning. The same conclusion also applies to the shifting operations.

It is meaningful for us to compare our proposed MT-LinAdapt algorithm with those discussed in the related work section. Different from the model adaptation based personalized sentiment classification solution proposed in [101], which treats the global model as fixed, MT-LinAdapt adapts the global model to capture the evolving nature of social norms. As a result, in [101] the individualized model adaptations are independent from each other; but in MT-LinAdapt, the individual learning tasks are coupled together to enable observation sharing across tasks, i.e., multi-task learning. Additionally, as illustrated in Eq (3.3.6) and (3.3.7), nonlinear model adaptation is achieved in MT-LinAdapt because of the different feature groupings in individual users and global model. This enables observations sharing across different feature groups, while in [101] observations can only be shared within the same feature group, i.e., linear model adaptation. Multi-task SVM introduced in [39] can be considered as a special case of MT-LinAdapt. In Multi-task SVM, only shifting operation is considered in individual users and the global model is simply estimated from the pooled observations across users.

Therefore, only linear model adaptation is achieved in Multi-task SVM and it cannot leverage prior knowledge conveyed in a predefined sentiment model.

3.4 Experimental Results

In this section, we perform empirical evaluations of the proposed MT-LinAdapt model. We verified the effectiveness of different feature groupings in individual users' and shared global model adaptation by comparing our solution with several state-of-the-art transfer learning and multi-task learning solutions for personalized sentiment classification, together with some qualitative studies to demonstrate how our model recognizes users' distinct expressions of sentiment.

3.4.1 Experimental Setup

• **Datesets.** We evaluated the proposed model on two large collections of review documents, i.e., Amazon product reviews [107] and Yelp restaurant reviews [108]. Each review document contains a set of attributes such as author ID, review ID, timestamp, textual content, and an opinion rating in discrete five-star range. We applied the following pre-processing steps on both datasets: 1) filtered duplicated reviews; 2) labeled reviews with overall rating above 3 stars as positive, below 3 stars as negative, and removed the rest; 3) removed reviewers who posted more than 1,000 reviews and those whose positive review ratio is more than 90% or less than 10% (little variance in their opinions and thus easy to classify). Since such users can be easily captured by the base model, the removal emphasizes comparisons on adapted models; 4) sorted each user's reviews in chronological order. Then, we performed feature selection by taking the union of top unigrams and bigrams ranked by Chi-square and information gain metrics [109], after removing a standard list of stopwords and porter stemming. The final controlled vocabulary consists of 5,000 and 3,071 textual features for Amazon and Yelp datasets respectively; and we adopted TF-IDF as the feature weighting scheme. From the resulting data sets, we randomly sampled 9,760 Amazon reviewers and 11,733 Yelp reviewers for testing purpose. There are 105,472 positive reviews and 37,674 negative reviews in the selected Amazon dataset; 108,105 positive reviews and 32,352 negative reviews in the selected Yelp dataset.

• **Baselines.** We compared the performance of MT-LinAdapt against seven different baselines, ranging from user-independent classifiers to several state-of-the-art model adaption methods and multi-task learning algorithms. Due to space limit, we will briefly discuss the baseline models below.

Our solution requires a user-independent classifier as base sentiment model for adaptation. We estimated logistic regression models from a separated collection of reviewers outside the preserved testing data on Amazon and Yelp datasets accordingly. We also included these isolated base models in our comparison and name them as *Base*. In order to verify the necessity of personalized sentiment models, we trained a global SVM based on the pooled adaptation data from all testing reviewers, and name it as *Global SVM*. We also estimated an independent SVM model for each single user only based on his/her adaptation reviews, and name it as *Individual SVM*. We included an instance-based transfer learning method [110], which considers the k-nearest neighbors of each testing review document from the isolated training set for personalized model training. As a result, for each testing case, we estimated an independent classification model, which is denoted as *ReTrain*. [111] used L2 regularization to enforce the adapted models to be close to the global model. We applied this method to get personalized logistic regression models and refer to it as *RegLR*. *LinAdapt* developed in [101] also performs group-wise linear model adaptation to build personalization classifiers. But it isolates model adaptation in individual users. *MT-SVM* is a multi-task learning method, which encodes task relatedness via a shared linear kernel [39].

• **Evaluation Settings.** We evaluated all the models with both synchronized (batch) and asynchronized (online) model update. We should note MT-SVM can only be tested in batch mode, because it is prohibitively expensive to retrain SVM repeatedly. In batch evaluation, we split each user's reviews into two sets: the first 50% for adaptation and the rest 50% for testing. In online evaluation, once we get a new testing instance, we first evaluate the up-to-date personalized classifier against the ground-truth; then use the instance to update the personalized model. To simulate the real-world situation where user reviews arrive sequentially and asynchronously, we ordered all reviews chronologically and accessed them one at a time for online model update. In particular, we utilized stochastic gradient descent for this online optimization [112]. Because of the biased class distribution in both datasets, we computed F1 measure for both positive and negative class in each user, and took macro average among users to compare the different models' performance. Both the source codes and data are available online [1].

[1] JNET. http://www.cs.virginia.edu/ lg5bt/mtlinadapt$_h tml/index.html$.

3.4.2 Effect of Feature Grouping

In MT-LinAdapt, different feature groupings can be postulated in individual users' and shared global model adaptation. Nonlinearity is introduced when different grouping functions are used in these two levels of model adaptation. Therefore, we first investigated the effect of feature grouping in MT-LinAdapt.

We adopted the feature grouping method named "*cross*" in [1] to cluster features into different groups. More specifically, we evenly spilt the training collection into N non-overlapping folds, and train a single SVM model on each fold. Then, we create a $V \times N$ matrix by putting the learned weights from N folds together, on which k-means clustering is applied to extract K feature groups. We compared the batch evaluation performance of varied combinations of feature groups in MT-LinAdapt. The experiment results are demonstrated in Table 4.1; and for comparison purpose, we also included the base classifier's performance in the table.

Table 3.1: Effect of different feature groupings in MT-LinAdapt.

Method	Amazon		Yelp	
	Pos F1	Neg F1	Pos F1	Neg F1
Base	0.8092	0.4871	0.7048	0.3495
400-800	0.8318	0.5047	0.8237	0.4807
400-1600	0.8385	0.5257	0.8309	0.4978
400-all	0.8441	0.5423	0.8345	0.5105
800-800	0.8335	0.5053	0.8245	0.4818
800-1600	0.8386	0.5250	0.8302	0.4962
800-all	0.8443	**0.5426**	**0.8361**	**0.5122**
1600-all	**0.8445**	0.5424	0.8357	0.5106
all-all	0.8438	0.5416	0.8361	0.5100

In Table 4.1, the two numbers in the first column denote the feature group sizes in personalized models and global model respectively. And *all* indicates one feature per group (i.e., no feature grouping). The adapted models in MT-LinAdapt achieved promising performance improvement against the base sentiment classifier, especially on the Yelp data set. As we increased the feature group size for global model, MT-LinAdapt's performance kept improving; while with the same feature grouping in the shared global model, a moderate size of feature groups in individual users is more advantageous.

These observations are expected. Because the global model is shared across users, all their adaptation reviews can be leveraged to adapt the global model so that sparsity is no longer an issue. Since more feature groups in the global model can be afforded, more accurate estimation of adaptation parameters can be achieved. But at the individual user level, data sparsity is still the bottleneck for

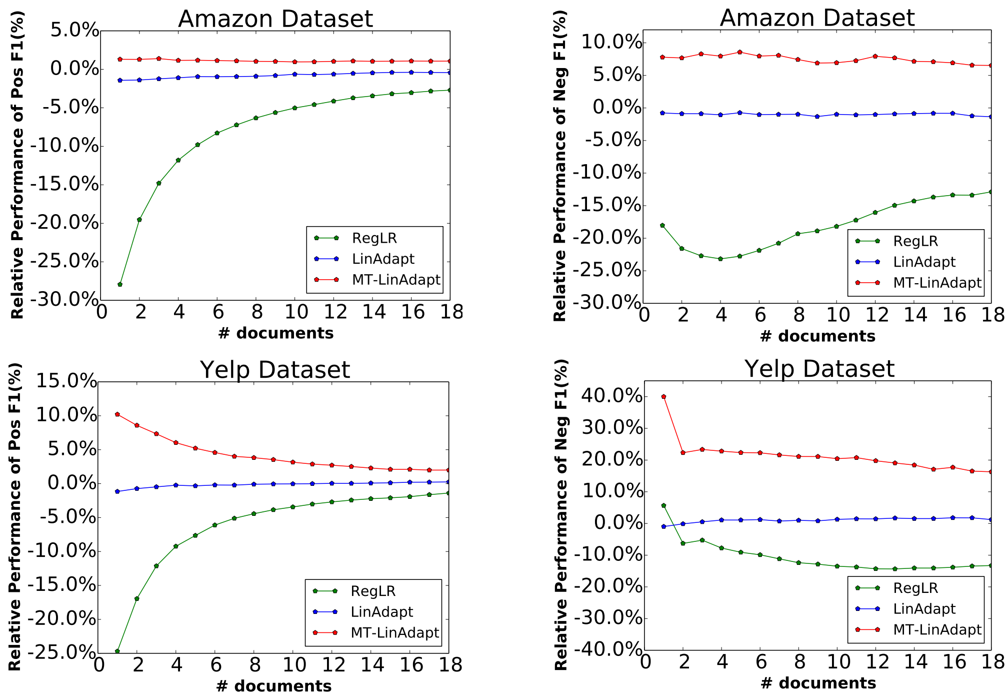

Figure 3.1: Relative performance gain between MT-LinAdapt and baselines on Amazon and Yelp datasets.

accurate adaptation estimation, and trade-off between observation sharing and estimation accuracy has to be made. Based on this analysis, we selected **800** and **all** feature groups for individual models and global model respectively in the following experiments.

3.4.3 Personalized Sentiment Classification

• **Synchronized model update.** Table 6.3 demonstrated the classification performance of MT-LinAdapt against all baselines on both Amazon and Yelp datasets, where binomial tests on win-loss comparison over individual users were performed between the best algorithm and runner-up to verify the significance of performance improvement. We can clearly notice that MT-LinAdapt significantly outperformed all baselines in negative class, and it was only slightly worse than MT-SVM on positive class. More specifically, per-user classifier estimation clearly failed to obtain a usable classifier, due to the sparse observations in single users. Model-adaptation based baselines, i.e., RegLR and LinAdapt, slightly improved over the base model. But because the adaptations across users are isolated and the base model is fixed, their improvement is very limited. As for negative class, MT-LinAdapt outperformed Global SVM significantly on both datesets. Since negative class suffers more from the biased prior distribution, the considerable performance improvement indicates effectiveness of our proposed personalized sentiment classification solution. As for positive class, the performance difference is not significant between MT-LinAdapt and MT-SVM on Amazon data set nor between MT-LinAdapt and Global SVM on Yelp data set. By looking into detailed results, we found that MT-LinAdapt outperformed MT-SVM on users with fewer adaptation reviews. Furthermore, though MT-SVM benefits from multi-task learning, it cannot leverage information from the given base classifier. Considering the biased class prior in these two data sets (2.8:1 on Amazon and 3.3:1 on Yelp), the improved classification performance on negative class from MT-LinAdapt is more encouraging.

Table 3.2: Classification results in batch mode.

Method	Amazon		Yelp	
	Pos F1	Neg F1	Pos F1	Neg F1
Base	0.8092	0.4871	0.7048	0.3495
Global SVM	0.8352	0.5403	**0.8411**	0.5007
Individual SVM	0.5582	0.2418	0.3515	0.3547
ReTrain	0.7843	0.4263	0.7807	0.3729
RegLR	0.8094	0.4896	0.7103	0.3566
LinAdapt	0.8091	0.4894	0.7107	0.3575
MT-SVM	**0.8484**	0.5367	0.8408	0.5079
MT-LinAdapt	0.8441	**0.5422***	0.8358	**0.5119***

$*$ indicates p-value<0.05 with Binomial test.

• **Asynchronized model update.** In online model estimation, classifiers can benefit from immediate update, which provides a feasible solution for timely sentiment analysis in large datasets. In this setting, only two baseline models are applicable without model reconstruction, i.e., RegLR and LinAdapt. To demonstrate the utility of online update in personalized sentiment models, we illustrate the relative performance gain of these models over the base sentiment model in Figure 3.1. The x-axis indicates the number of adaptation instances consumed in online update from all users, i.e., the 1st review means after collecting the first review of each user.

MT-LinAdapt converged to satisfactory performance with only a handful of observations in each user. LinAdapt also quickly converged, but its performance was very close to the base model, since no observation is shared across users. RegLR needs the most observations to estimate satisfactory personalized models. The improvement in MT-LinAdapt demonstrates the benefit of shared model adaptation, which is vital when the individuals' adaptation data are not immediately available but timely sentiment classification is required.

It is meaningful to investigate how the shared global model and personalized models are updated during online learning. The shift in the shared global model reflects changes in social norms, and the discrepancy between the shared global model and personalized models indicates the variances of

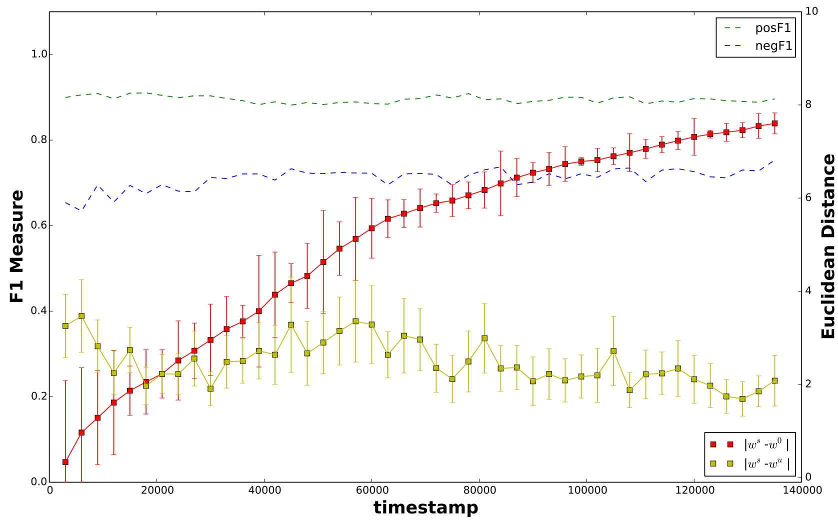

Figure 3.2: Online model update trace on Amazon.

individuals' opinions. In particular, we calculated Euclidean distance between global model w^s and base model w^0 and that between individualized model w^u and shared global model w^s during online model updating. To visualize the results, we computed and plotted the average Euclidean distances in every 3000 observations during online learning, together with the corresponding variance. To illustrate a comprehensive picture of online model update, we also plotted the corresponding average F1 performance for both positive and negative class. Because the Euclidean distance between w^s and w^0 is much larger than that between w^s and w^u, we scaled $||w^s - w^0||$ by 0.02 on Amazon dataset in Figure 3.2. Similar results were observed on Yelp data as well; but due to space limit, we do not include them.

As we can clearly observe that the difference between the base model and newly adapted global model kept increasing during online update. At the earlier stage, it is increasing much faster than the later stage, and the corresponding classification performance improves more rapidly (especially in negative class). The considerably large variance between w^0 and w^s at the beginning indicates the divergence between old and new social norms across users. Later on, variance decreased and converged with more observations, which can be understood as the formation of the new social norms among users. On the other hand, the distance between personalized models and shared global model fluctuated a lot at the beginning; with more observations, it became stable later on. This is also reflected in the range of variance: the variance is much smaller in later stage than earlier stage, which indicates users comply to the newly established social norms.

Table 3.3: Shared model adaptation for cold start on Amazon and Yelp.

Obs.	Amazon						Yelp					
	Shared-SVM		MT-SVM		MT-LinAdapt		Shared-SVM		MT-SVM		MT-LinAdapt	
	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}	0.9004	0.7013	**0.9264**	0.7489	0.9122	**0.7598**	0.7882	0.5537	**0.9040**	0.7201	0.8809	**0.7306**
2^{nd}	0.9200	0.6872	**0.9200**	**0.7319**	0.8945	0.7292	0.7702	0.5266	**0.8962**	0.6959	0.8598	**0.6968**
3^{rd}	0.9164	0.6967	**0.9164**	0.7144	0.8967	**0.7260**	0.7868	0.5278	**0.9063**	**0.7099**	0.8708	0.7069

3.4.4 Shared Adaptation Against Cold Start

Cold start refers to the challenge that a statistic model cannot draw any inference for users before sufficient observations are gathered [113]. The shared model adaptation in MT-LinAdapt helps alleviate cold start in personalized sentiment analysis, while individualized model adaptation method, such as RegLR and LinAdapt, cannot achieve so. To verify this aspect, we separated both Amazon and Yelp reviewers into two sets: we randomly selected 1,000 reviewers from the isolated training set and exhausted all their reviews to estimate a shared SVM model, MT-LinAdapt and MT-SVM. Then those models were directly applied onto the testing reviewers for evaluation. Again, because it is time consuming to retrain a SVM model repeatedly, only MT-LinAdapt performed online model update in this evaluation. We report the performance on the first three observations from all testing users accordingly in Table 3.3.

MT-LinAdapt achieved promising performance on the first testing cases, especially on the negative class. This indicates its estimated global model is more accurate on the new testing users. Because MT-SVM cannot be updated during this online test, only its previously estimated global model from the 1,000 training users can be applied here. As we can notice, its performance is very similar to the shared SVM model (especially on Amazon dataset). MT-LinAdapt adapts to this new collection of users very quickly, so that improved performance against the static models at later stage is achieved.

3.4.5 Vocabulary Stability

One derivative motivation for personalized sentiment analysis is to study the diverse use of vocabulary across individual users. We analyzed the variance of words' sentiment polarities estimated in the personalized models against the base model. Table 3.4 shows the most and the least variable features on both datasets. It is interesting to find that words with strong sentiment polarities tend to be more stable across users, such as "disgust," "regret," and "excel." This demonstrates the sign of

Table 3.4: Top six words with the highest and lowest variances of learned polarities by MT-LinAdapt.

Amazon	Highest	cheat astound	healthi the-wrong	enjoy-read the-amaz
	Lowest	mistak regret	favor perfect-for	excel great
Yelp	Highest	total-worth advis	lazi impress	was-yummi so-friend
	Lowest	omg frustrat	veri-good disgust	hungri a-must

conformation to social norms. There are also words exhibiting high variances in sentiment polarity, such as "was-yummi," "lazi," and "cheat," which indicates the heterogeneity of users' opinionated expressions.

3.5 Conclusion

In this work, we captured users' diverse preferences in expressing attitudes via personalized sentiment classification. In particularly, it is achieved by the notion of shared model adaptation, which is motivated by the social theories that humans' opinions are diverse but shaped by the ever-changing social norms. In the proposed MT-LinAdapt algorithm, global model sharing alleviates data sparsity issue, and individualized model adaptation captures the heterogeneity in humans' sentiments and enables efficient online model learning. Extensive experiments on two large review collections from Amazon and Yelp confirmed the effectiveness of our proposed solution.

The information sharing is achieved via multi-task learning by treating each individual user as a single task, which can be further enhanced to alleviate the data sparsity issue, i.e., via the grouping of similar users. The user groups can be automatically identified to maximize the effectiveness of shared model adaptation. Users in the same group share the same model parameters by contributing all their observations for optimizing group-wise model parameters.

Chapter 4

Clustered Model Adaptation for Personalized Sentiment Analysis

In this Chapter, we propose to capture humans' variable and idiosyncratic ways of expressing attitudes 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 exploit the clustering property of users' opinions via imposing a non-parametric Dirichlet Process prior over the personalized models. 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.

4.1 Introduction

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 [31] 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 [32] 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 [33], 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 [34]. 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' 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.

Due to the relatedness among the tasks, the proposed solution can also be understood from the perspective of multi-task learning [39, 40, 89]. 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 [101]. 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.

4.2 Related work

The proposed method utilizes the relatedness among users to perform user grouping, thus is closely related with clustering-based user modeling algorithms. [114] 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 [80, 115]. Dirichlet Process prior [32] naturally satisfies this goal: it associates related tasks into groups via exploiting the clustering property of data. [116] 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 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. [40, 117] estimated a set of linear classifiers in automatically identified groups. However, due to the sparsity of personal opinionated data in the sentiment analysis, a full set of model parameters have to be estimated in each task. Our solution instead only learns simple model transformations over groups of features in each user group [101], 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.

4.3 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.

4.3.1 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 [118]. 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. [119] 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 [33, 120] 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 [96]. 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.

4.3.2 Personalized Model Adaptation

Following the assumptions described in Section 3.3.2, we also 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. 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 $D^u = \{(\boldsymbol{x}_d^u, y_d^u)\}_{d=1}^{|D^u|}$. Each document d is represented by a V-dimension vector $\boldsymbol{x}_d$ of textual features, and y_d is the corresponding sentiment label. We assume each user is associated with a sentiment model $f(\boldsymbol{x}; \boldsymbol{w}^u) \rightarrow y$, which is characterized by the individualized feature weight vector $\boldsymbol{w}^u$. Estimating $f(\boldsymbol{x}; \boldsymbol{w}^u)$ for users in U is the inductive learning task of our focus. Borrowing the techniques of linear transformations, we assume each user's personalized model is obtained from a global sentiment model $f(\boldsymbol{x}; \boldsymbol{w}^s)$ via a series of linear model transformations [1, 101], i.e., shift and scale the shared model parameter $\boldsymbol{w}^s$ into $\boldsymbol{w}^u$ based on D^u. To simplify the discussions in this paper, we also assume binary sentiment classification, i.e., $y \in \{0, 1\}$, and we 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 [1], which is also utilized and introduced in Section 3.3.2. By defining $g(i) \rightarrow k$ 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 $\boldsymbol{\theta}^u = (a_1^u, a_2^u, \ldots, a_K^u, b_1^u, b_2^u, \ldots, b_K^u)$, where a_k^u and b_k^u 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 $\boldsymbol{w}^s$ to personalized model $\boldsymbol{w}^u$ as $\forall i \in \{1, 2, \ldots, V\}$, $\boldsymbol{w}_i^u = a_{g(i)}^u \boldsymbol{w}_i^s + b_{g(i)}^u$. Because $\boldsymbol{\theta}^u$ uniquely determines the personalized feature weight vector $\boldsymbol{w}^u$, we will then refer to $\boldsymbol{\theta}^u$ as the personalized sentiment model for user u in our discussions.

Different from what has been explored in [1, 101], where the global model $\boldsymbol{w}^s$ is predefined and fixed, we assume $\boldsymbol{w}^s$ 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 $\boldsymbol{w}^s$ from a predefined sentiment model $\boldsymbol{w}^0$. $\boldsymbol{w}^0$ 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 $\boldsymbol{\theta}^s = (a_1^s, a_2^s, \dots, a_L^s, b_1^s, b_2^s, \dots, b_L^s)$, 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 $\boldsymbol{w}^0$ can serve as a prior for global sentiment classification [101]. 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,

$$P(y_d^u = 1 | \boldsymbol{x}_d^u, \boldsymbol{\theta}^u, \boldsymbol{\theta}^s, \boldsymbol{w}^0) = \sigma\Big(\sum_{k=1}^{K} \sum_{g(i)=k} (a_k^u \boldsymbol{w}_i^s + b_k^u) \boldsymbol{x}_{d,i}^u\Big) \tag{4.3.1}$$

where $\boldsymbol{w}_i^s = a_{g'(i)}^s \boldsymbol{w}_i^0 + b_{g'(i)}^s$ and $\sigma(x) = \frac{1}{1+\exp(-\boldsymbol{x})}$.

4.3.3 Non-parametric Modeling of Groups

The inductive learning task in each user u hence becomes to estimate $\boldsymbol{\theta}^u$ that maximizes the likelihood of the user's own opinionated data defined by Eq (5.3.1). Accordingly, a shared task for all users is to estimate $\boldsymbol{\theta}^s$ 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 $\{\boldsymbol{\theta}^u\}_{u=1}^N$ 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 [121] or individually trained models [122], 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; 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 $\{\boldsymbol{\theta}^u\}_{u=1}^N$ as fixed but unknown, we treat it as stochastic by assuming each user's model parameter $\boldsymbol{\theta}^u$ is drawn from a Dirichlet Process prior [32, 123]. 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,

$$\begin{aligned} G &\sim DP(\alpha, G_0), \\ \boldsymbol{\theta}^u|G &\sim G, \\ y_d^u|\boldsymbol{x}_d^u, \boldsymbol{\theta}^u, \boldsymbol{\theta}^s, \boldsymbol{w}^0 &\sim P(y_d^u = 1|\boldsymbol{x}_d^u, \boldsymbol{\theta}^u, \boldsymbol{\theta}^s, \boldsymbol{w}^0). \end{aligned} \tag{4.3.2}$$

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 $\boldsymbol{\theta}^u$. To simplify the notations for discussion, we define $\boldsymbol{a}^u$ and $\boldsymbol{b}^u$ as the scaling and shifting components in $\boldsymbol{\theta}^u$, such that $\boldsymbol{\theta}^u = (\boldsymbol{a}^u, \boldsymbol{b}^u)$. We impose an isometric Gaussian distribution in G_0 over $\boldsymbol{\theta}^u$ as $\boldsymbol{\theta}^u \sim N(\mu, \sigma^2)$, where $\mu = (\mu^a, \mu^b)$ and $\sigma = (\sigma^a, \sigma^b)$ 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 $\boldsymbol{\theta}^s$ as a latent random variable, and impose another isometric Gaussian prior over it as $\boldsymbol{\theta}^s \sim N(\mu_s, \sigma_s)^2$, where μ_s and σ_s are also decomposed with respect to the shifting and scaling operations.

By integrating out G in Eq (4.3.2), the predictive distribution of $\boldsymbol{\theta}^u$ conditioned on the individualized models in the other users, denoted as $\boldsymbol{\theta}^{-u} = \{\boldsymbol{\theta}^1, .., \boldsymbol{\theta}^{u-1}, \boldsymbol{\theta}^{u+1}, ...\boldsymbol{\theta}^N\}$, can be analytically computed as follows,

$$p(\boldsymbol{\theta}^u|\boldsymbol{\theta}^{-u}, \alpha, G_0) = \frac{\alpha}{N-1+\alpha} G_0 + \frac{1}{N-1+\alpha} \sum_{j \neq i}^{N} \delta_{\theta^u}(\boldsymbol{\theta}^j) \tag{4.3.3}$$

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

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 $\boldsymbol{\theta}^j = \boldsymbol{\theta}^u$, 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 (4.3.3) 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 $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \dots, \boldsymbol{\phi}_c\}$, i.e., the group models, where the group index c takes value from 1 to ∞, and $\boldsymbol{\phi}_i$ 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 (4.3.2) only has finite supports at the points of $\{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \dots, \boldsymbol{\theta}^N\}$, 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 [124] as follows:

$$\begin{aligned} &P(y^u|\boldsymbol{x}^u, \boldsymbol{w}^0, \alpha, G_0) \qquad (4.3.4) \\ &= \int d\boldsymbol{\phi} \int d\boldsymbol{\theta}^s \int d\boldsymbol{\pi} \sum_{c_u=1}^{\infty} \prod_{d=1}^{|D^u|} P(y^u_d|\boldsymbol{x}^u_d, \boldsymbol{\phi}_{c_u}, \boldsymbol{\theta}^s, \boldsymbol{w}^0) p(c_u|\boldsymbol{\pi}) p(\boldsymbol{\phi}_{c_u}|\mu, \sigma^2) p(\boldsymbol{\theta}^s|\mu_s, \sigma_s^2) p(\boldsymbol{\pi}|\alpha) \end{aligned}$$

where $\boldsymbol{\pi} = (\pi_c)_{c=1}^{\infty} \sim Stick(\alpha)$ captures the proportion of unique sample $\boldsymbol{\phi}_c$ in the whole collection. And the stick-breaking process $Stick(\alpha)$ for $\boldsymbol{\pi}$ 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., $\{\boldsymbol{\phi}_c\}_{c=1}^{\infty}$) is a set of linear model transformations, we name the resulting model defined by Eq (4.3.4) 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 4.1. We should note that our cLinAdapt model is not a fully generative model: as defined in Eq (4.3.4), we treat the documents

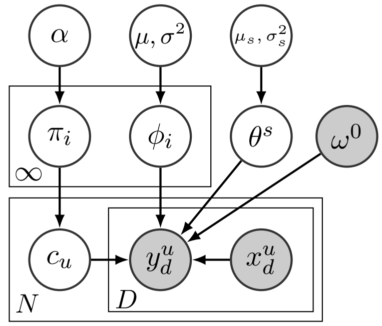

Figure 4.1: Graphical model representation of cLinAdapt. Light circles denote the latent random variables, and shadow circles denote the observed ones. The outer plate indexed by N denotes the users in the collection, the inner plate indexed by D denotes the observed opinionated data associated with user u, and the upper plate denotes the parameters for the countably infinite number of latent user groups in the collection.

$\{\boldsymbol{x}^u\}_{u=1}^N$ 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.

4.4 Posterior Inference

To apply cLinAdpat for personalized sentiment classification, we need to infer the posterior distributions of: 1) group-wise model adaptation parameters $\{\boldsymbol{\phi}_c\}_{c=1}^\infty$, each one of which captures the homogeneity of personalized sentiment models in a corresponding latent user group; 2) global model adaptation parameter $\boldsymbol{\theta}^s$, which is shared by all users' sentiment models; 3) group membership variable c_u for user u; and 4) sentiment labels 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) [125] 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 $\{\boldsymbol{\phi}_c\}_{c=1}^\infty$ and global model $\boldsymbol{\theta}^s$, and then maximum likelihood estimation for $\{\boldsymbol{\phi}_c\}_{c=1}^\infty$ and $\boldsymbol{\theta}^s$ 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 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 [126], we introduce a set of auxiliary random variables of size m, i.e., $\{\boldsymbol{\phi}_i^a\}_{i=1}^m$, 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,

$$P\big(c_u = c|y^u, \boldsymbol{x}^u, \{\boldsymbol{\phi}_i\}_{i=1}^C, \{\boldsymbol{\phi}_j^a\}_{j=1}^m, \boldsymbol{\theta}^s, \boldsymbol{w}^0\big) \propto \begin{cases} n_c^{-u} \prod_{d=1}^{|D^u|} P(y_d^u|\boldsymbol{x}_d^u, \boldsymbol{\phi}_c, \boldsymbol{\theta}^s, \boldsymbol{w}^0) & \text{for} \quad 1 \le c \le C, \\ \frac{\alpha}{m} \prod_{d=1}^{|D^u|} P(y_d^u|\boldsymbol{x}_d^u, \boldsymbol{\phi}_c^a, \boldsymbol{\theta}^s, \boldsymbol{w}^0) & \text{for} \quad 1 < c \le m. \end{cases} \tag{4.4.1}$$

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

Because of the introduction of auxiliary variables $\{\boldsymbol{\phi}_i^a\}_{i=1}^m$, the integration of $\{\boldsymbol{\phi}_c\}_{c=1}^\infty$ 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 $\boldsymbol{\theta}^u$ in G_0 are Gaussian, sampling the auxiliary variables is efficient.

We should note that the sampling step derived in Eq (4.4.1) 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 $\{\boldsymbol{\phi}_c\}_{c=1}^{\infty}$ and $\boldsymbol{\theta}^s$:** 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 $\{\boldsymbol{\phi}_c\}_{c=1}^{\infty}$ and $\boldsymbol{\theta}^s$ 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 $\{\boldsymbol{\phi}_c\}_{c=1}^{C}$ and $\boldsymbol{\theta}^s$ can be written as,

$$L\big(\{\boldsymbol{\phi}_c\}_{c=1}^{C}, \boldsymbol{\theta}^s\big) = \sum_{u=1}^{N} \log P(y^u|\boldsymbol{x}^u, \boldsymbol{\phi}_{c_u}, \boldsymbol{\theta}^s, \boldsymbol{w}^0) + \sum_{c=1}^{C} \log p(\boldsymbol{\phi}_c|\mu, \sigma^2) + \log p(\boldsymbol{\theta}^s|\mu_s, \sigma_s^2) \tag{4.4.2}$$

As the global model adaptation parameter $\boldsymbol{\theta}^s$ is shared by all the users (as defined in Eq (5.3.1)), it makes the estimation of $\{\boldsymbol{\phi}_c\}_{c=1}^{C}$ dependent across all the user groups, i.e., information sharing across groups in cLinAdapt.

In Section 4.3.3, we did not specify the dconfiguration of the prior distributions on $\boldsymbol{\theta}^u$ and $\boldsymbol{\theta}^s$, i.e., Gaussian's mean and standard deviation. But given $\boldsymbol{\theta}^u$ and $\boldsymbol{\theta}^s$ 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., $\mu^a = 1$ and $\mu^b = 0$, 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 $\boldsymbol{\theta}^s$.

Eq (5.3.13) can be efficiently maximized by a gradient-based optimizer, and the actual gradients of Eq (5.3.13) 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 $\boldsymbol{\phi}_c$ and $\boldsymbol{\theta}^s$ on a specific training instance $(\boldsymbol{x}_d^u, y_d^u)$ in user u:

$$\frac{\partial L(\cdot)}{\partial a_k^{c_u}} = \Delta_d^u \sum_{g(i)=k} \Big(a_{g'(i)}^s \boldsymbol{w}_i^0 + b_{g'(i)}^s\Big) \boldsymbol{x}_{di}^u + \frac{a_k^{c_u} - 1}{\sigma^2} \tag{4.4.3}$$

$$\frac{\partial L(\cdot)}{\partial a_l^s} = \Delta_d^u \sum_{g'(i)=l} a_{g(i)}^{c_u} \boldsymbol{w}_i^0 \boldsymbol{x}_{di}^u + \frac{a_l^s - 1}{\sigma_s^2} \tag{4.4.4}$$

where $\Delta_d^u = y_d^u - P(y_d^u = 1|\boldsymbol{x}_d^u, \boldsymbol{\phi}_{c_u}, \boldsymbol{\theta}^s, \boldsymbol{w}^0)$, 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 $\boldsymbol{\phi}_c$ (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 (4.4.4)), which adds another dimension of task relatedness for multi-task learning. Also as illustrated in Eq (4.4.3) and (4.4.4), when different feature groupings are used in $g(\cdot)$ and $g'(\cdot)$, nonlinearity is introduced to propogate information across features.

• **Predict for 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 y^u in user u's testing documents by,

$$P(y_d^u = 1 | \boldsymbol{x}_d^u, \{\boldsymbol{\phi}_c^t\}_{c=1}^{C_t}, \boldsymbol{\theta}_t^s, \boldsymbol{w}^0) = \sum_{c=1}^{C_t} P(c_u^t = c) P(y_d^u = 1 | \boldsymbol{x}_d^u, \boldsymbol{\phi}_{c_u^t}^t, \boldsymbol{\theta}_t^s, \boldsymbol{w}^0) \tag{4.4.5}$$

where $\left(\{\boldsymbol{\phi}_c^t\}_{c=1}^{C_t}, c_u^t, \boldsymbol{\theta}_t^s\right)$ are the estimates of latent variables at the tth iteration, $P(c_u^t = c)$ is estimated in Eq (4.4.1) and $P(y_d^u = 1 | x_d^u, \phi_{c_u^t}, \theta^s, w^0)$ is computed by Eq (5.3.1). Then the posterior of $\boldsymbol{y}^u$ can thus be estimated via empirical expectation after T iterations,

$$P(y_d^u = 1 | \boldsymbol{x}_d^u, \boldsymbol{w}^0, \alpha, G_0) = \frac{1}{T} \sum_{t=1}^{T} P(y_d^u = 1 | \boldsymbol{x}_d^u, \{\boldsymbol{\phi}_c^t\}_{c=1}^{C_t}, \boldsymbol{\theta}_t^s, \boldsymbol{w}^0)$$

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.

4.5 Experimental Results 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.

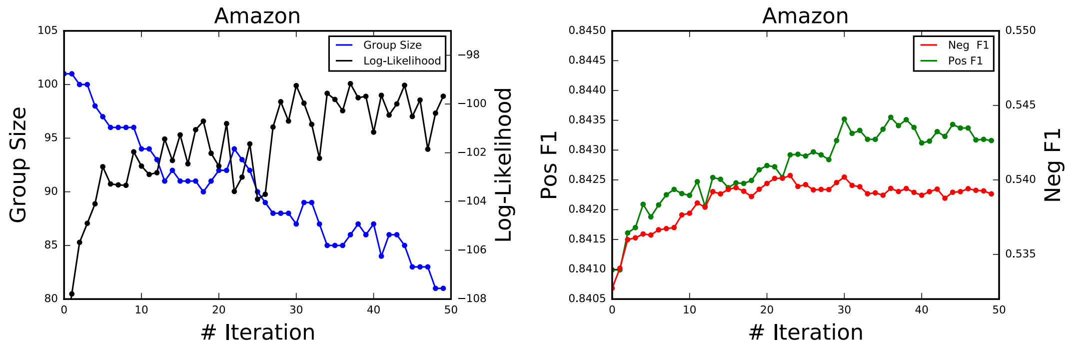

Figure 4.2: Trace of likelihood, group size and performance during iterative posterior sampling in cLinAdapt for Amazon.

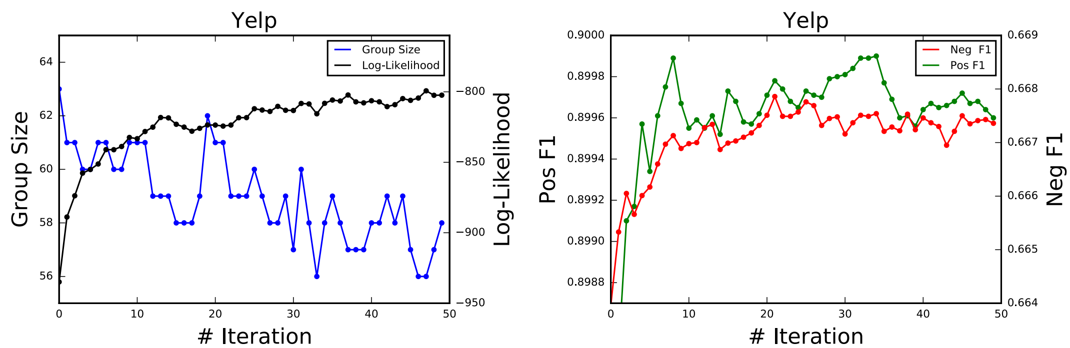

Figure 4.3: Trace of likelihood, group size and performance during iterative posterior sampling in cLinAdapt for Yelp.

4.5.1 Experimental Setup

• **Datesets.** We used the same review datasets as in Section 3.4, Amazon [107] 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. This sparsity in the two datasets raises a serious challenge for personalized sentiment analysis.

We directly used the same processed Amazon data as in Section 3.4 in the following evaluation. While for Yelp dataset, they update their data every 6 months. Thus, we get the latest data and performed the same pre-processing steps: 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. From the resulting data, we randomly sampled 10,830 Yelp reviewers for evaluation purpose. The 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. There are 105,472

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

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., w^0 in Eq (5.3.1)). 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 [101]. 5) **LinAdapt+kMeans**: To verify the effectiveness of our proposed user grouping method in personalized sentiment model learning, we followed [121] 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 [111] 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 [40]. 8) **MT-SVM**: It is a state-of-the-art multi-task learning solution proposed in [39]. 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 4.3.3 and 4.4, 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. Both the source codes and data are available online [2].

4.5.2 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 (5.3.3)) during each iteration of posterior inference in cLinAdapt over all users from both datasets. We reported the results for the two datasets in Figure 4.2 and 4.3, 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%

[2] JNET. http://www.cs.virginia.edu/ lg5bt/clinadapt$_h tml/index.html$.

Figure 4.4: Word clouds on Amazon (left) and Yelp (right).

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.

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 4.4 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.

Table 4.1: 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

4.5.3 Effect of Feature Grouping

We then investigated the effect of feature grouping in cLinAdapt. As discussed in Section 4.3.3, 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 [1]. 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 4.1; and for comparison purpose, we also included the base classifier's performance in the table. In Table 4.1, 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.

4.5.4 Personalized Sentiment Classification

We compared cLinAdapt against all nine baselines on both Amazon and Yelp datasets, and the detailed performance is reported in Table 6.3. 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

Table 4.2: 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

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.

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 (4.4.1) 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

Table 4.3: Effectiveness of model sharing for cold-start on Amazon.

Obs.	Individual SVM		LinAdapt		MT-SVM		cLinAdapt	
	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**
2^{nd}	0.4683	0.3831	0.8455	0.5495	0.8534	0.5267	**0.8795**	**0.6076**
3^{rd}	0.7362	0.1751	0.8113	0.4863	0.8283	0.4919	**0.8440**	**0.5402**

Table 4.4: Effectiveness of model sharing for cold-start on Yelp.

Obs.	Individual SVM		LinAdapt		MT-SVM		cLinAdapt	
	Pos F1	Neg F1	Pos F1	Neg F1	Pos F1	Neg F1	Pos F1	Neg F1
1^{st}	0.0000	0.4101	0.9322	0.7724	0.9251	0.7285	**0.9582**	**0.8335**
2^{nd}	0.7402	0.3116	0.9243	0.7176	0.9291	0.7027	**0.9501**	**0.7726**
3^{rd}	0.7812	0.1608	0.8873	0.6639	0.8954	0.6619	**0.9116**	**0.7147**

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 4.4.

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.

Another observation in Table 4.4 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 6.3. 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.

4.6 Conclusion

In this work, we developed a clustered model adaptation solution for understanding users' diverse preferences in expressing opinions. 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.

The relatedness among users are encoded implicitly by the task relatedness, i.e., multi-task learning. While the available network structure provides explicit relatedness among users, thus it is interesting to study whether we can gain a comprehensive understanding of user intents by incorporating network structure and text content.

Part II

Incorporating Network for Holistic User Behavior Modeling

Chapter 5

A Holistic User Behavior Modeling via Multi-task Learning

In this Chapter, we perform user modeling to understand user intents by utilizing multiple modalities. It is critical to achieve this goal 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.

5.1 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 [6, 7]. Numerous successes have proved its value in practical applications. For example, Yan et al. [127] 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. [3] found that modeling users' review content 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 [28, 43]. 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 [7] or their decision making process [3, 6]. 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 [22], and network structure has been analyzed to examine users' social grouping and belonging [35–37].

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*** [38] 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 content 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., a multi-task learning approach [39–41]. In the first task, a statistical language model is used to model the generation of textual content, and a logistic regression model maps the textual content to the sentiment polarity. In the second task, a stochastic block model [128] 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 [126] 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 [129]. 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) [130] 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*** [131] 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.

5.2 Related Work

A lof of efforts are devoted to combine text content with network structure to improve the fidelity of learned user models. Speriosu et al. [74] proposed to propagate labels from a supervised classifier over the Twitter follower graph to improve sentiment classification. Studies in [23, 75] incorporated user-user interactions as side information to regularize sentiment classification. Cheng et al. [76] leveraged signed social network to infer the sentiment of text documents in an unsupervised manner. Tang et al. [77] 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. Pozzi et al. [78] utilized the approval relations to estimate user polarities about a given topic in a semi-supervised framework.

5.3 Methodology

5.3.1 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 $D_i = \big\{(\boldsymbol{x}_d^i, y_d^i)\big\}_{d=1}^{|D_i|}$. 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 content 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.

5.3.2 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 $(\boldsymbol{x}_d^i, y_d^i)$. 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 content to binary sentiment polarities as,

$$p(y_d^i|\boldsymbol{x}_d^i, z_d^i) = \frac{1}{1+\exp(-y_d^i \bar{\boldsymbol{\phi}}_{\boldsymbol{z_d^i}}^{\mathsf{T}} \boldsymbol{x}_d^i)}, \tag{5.3.1}$$

where $\bar{\boldsymbol{\phi}}_{\boldsymbol{z_d^i}}$ is a V-dimensional feature weight vector. Following the setting in [41] 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 $\boldsymbol{\phi_s}$ shared by all models, and one local component $\boldsymbol{\phi_c}$ just for this current model, we further decompose $\bar{\boldsymbol{\phi}}_{\boldsymbol{c}} = \boldsymbol{\phi_s} + \boldsymbol{\phi_c}$ in our logistic regression model.

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

$$p(D_i|\boldsymbol{\pi}_i) = \prod_{d \in D_i} \sum_{z_d^i=1}^{C} p(y_d^i|\boldsymbol{x}_d^i, z_d^i) p(\boldsymbol{x}_d^i|z_d^i) p(z_d^i|\boldsymbol{\pi}_i) \tag{5.3.2}$$

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 [128] 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,

$$p(E_i|\boldsymbol{\pi}_i, z_{j\to i}, B) = \prod_{e_{ij} \in E_i} \sum_{z_{i \to j}=1}^{C} p(e_{ij}|B_{z_{i\to j}, z_{j\to i}}) p(z_{i\to j}|\boldsymbol{\pi}_i) \tag{5.3.3}$$

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

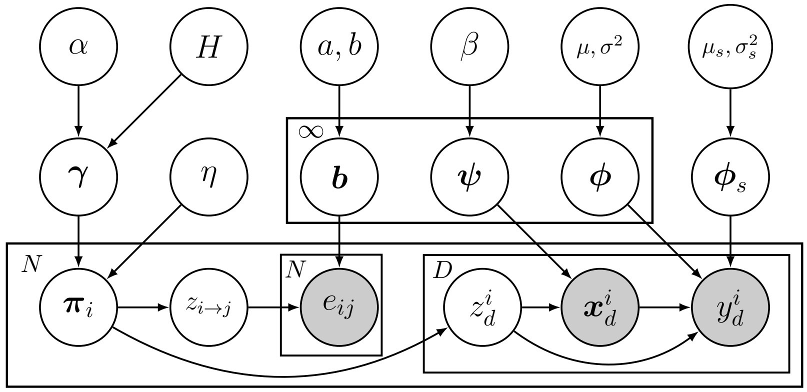

Figure 5.1: Graphical model representation of HUB. The upper plate indexed by ∞ denotes the unified model parameters for collective identities. The outer plate indexed by N denotes distinct users. The inner plates indexed by N and D denote each user's social connections and review documents respectively.

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 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 [32] 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:

$$p(D, E) = \sum_{c=1}^{\infty} \gamma_c \delta_{\boldsymbol{\theta}_c}, \tag{5.3.4}$$

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 $N(\mu, \sigma^2)$ 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 $N(\mu_s, \sigma_s^2)$ over the globally shared logistic regression parameter $\boldsymbol{\phi}_s$.

The global mixture structure defined in Eq (5.3.4) 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 (5.3.3)). 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 6.1, using a graphical model representation.

5.3.3 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 content under each collective identity; $\boldsymbol{\phi}_c$ and $\boldsymbol{\phi}_s$ capture the mapping from textual content 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 $z_{i\rightarrow j}$ 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) [125] based iterative algorithm for posterior inference in these three types of latent variables. More specifically, Gibbs Sampling method based on auxiliary variables [126] 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., $\{z_{i\rightarrow j}\}_{j\neq i}^N$. 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 $z_{i\rightarrow j}$ 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,

$$p(\boldsymbol{\pi}_i, z_d^i|D_i, E_i, \boldsymbol{\gamma}, \alpha, \eta, \Psi, \Phi) \propto p(\{z_d^i\}_{d=1}^D|\boldsymbol{\pi}_i)p(\{z_{i\rightarrow j}\}_{j\neq i}^N|\boldsymbol{\pi}_i)p(\boldsymbol{\pi}_i|\boldsymbol{\gamma}, \eta)p(y_d^i, x_d^i|z_d^{u_i}, \boldsymbol{\psi}_{z_d^i}, \boldsymbol{\phi}_{z_d^i}). \tag{5.3.5}$$

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 $p(\{z_{i\rightarrow j}\}_{j\neq i}^N|\boldsymbol{\pi}_i)$, we can marginalize out $\boldsymbol{\pi}_i$ in Eq (5.3.5). This leaves us the conditional probability of z_d^i in user u_i given his/her rest collective identity assignments,

$$p(z_d^i|\boldsymbol{\gamma}, \eta) = \frac{\Gamma(\eta)}{\Gamma(\eta + n_{i\star} + l_{i\star\star})} \prod_{c=1}^C \frac{\Gamma(\eta\boldsymbol{\gamma}_c + n_{i,c} + l_{i\star,c})}{\Gamma(\eta\gamma_c)}, \tag{5.3.6}$$

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 (5.3.5) can be computed as follows:

$$p(z_d^i = c|D_i, E_i, \boldsymbol{\gamma}, \alpha, \eta, \Psi, \Phi) \propto (n_{i,c}^{-d} + l_{i\star,c} + \eta\gamma_c)p(y_d^i, x_d^i|\boldsymbol{\psi}_c, \boldsymbol{\phi}_c), \tag{5.3.7}$$

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 [126] 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,

$$\begin{aligned}
&p\big(z_d^i = c|D_i, E_i, \Phi, \{\boldsymbol{\phi}_m^a\}_{m=1}^M, \Psi\big) \qquad (5.3.8)\\
&\propto (n_{i,c}^{-d} + l_{i\star,c} + \eta\gamma_c) p(y_d^i, x_d^i|\boldsymbol{\psi}_c, \boldsymbol{\phi}_c)\\
&= \begin{cases} (n_{i,c}^{-i,d} + l_{i\star,c} + \eta\boldsymbol{\gamma}_c) p(y_d^i, x_d^i|\boldsymbol{\psi}_c, \boldsymbol{\phi}_c) & \text{for } 1 \le c \le C,\\ \frac{\eta\gamma_e}{M} p(y_d^i, x_d^i|\boldsymbol{\psi}_c, \boldsymbol{\phi}_c^a) & \text{for } C < c \le C+M. \end{cases}
\end{aligned}$$

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,

$$\begin{aligned}
&p(y_d^i, \boldsymbol{x}_d^i|\boldsymbol{\psi}_c, \boldsymbol{\phi}_c) = \qquad (5.3.9)\\
&\begin{cases} \frac{1}{1+\exp(-y_d^i \bar{\boldsymbol{\phi}}_c^T \boldsymbol{x}_d^i)} \frac{m_{d\star}^i!}{\prod_{v=1}^V m_{d,v}^i!} \prod_{v=1}^V {\psi_{c,v}}^{m_{d,v}^i} & \text{for } 1 \le c \le C,\\ \frac{1}{1+\exp(-y_d^i {\bar{\boldsymbol{\phi}}_c^{\boldsymbol{a}}}^T \boldsymbol{x}_d^i)} \frac{\Gamma(\beta)}{\Gamma(\beta + m_{d\star}^i)} \prod_{v=1}^V \frac{\Gamma(m_{d,v}^i + \beta_v)}{\Gamma(\beta_v)} & \text{for } C < c \le C+M \end{cases}
\end{aligned}$$

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 [41], 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,

$$p(\boldsymbol{\pi}_i, z_{i\to j}|D_i, E_i, \boldsymbol{\gamma}, \alpha, \eta, \Psi, \Phi) \propto p(\{z_d^i\}_{d=1}^D|\boldsymbol{\pi}_i) \\ p(\{z_{i\to j}\}_{j\neq i}^N|\boldsymbol{\pi}_i)p(\boldsymbol{\pi}_i|\boldsymbol{\gamma}, \eta)p(e_{ij}|z_{i\to j}, z_{j\to i}, B) \tag{5.3.10}$$

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,

$$p(z_{i\to j} = c|D_i, E_i, \boldsymbol{\gamma}, \alpha, \eta, \Psi, \Phi) \propto (n_{i,c} + l_{i\star,c}^{-(i\to j)} + \eta\gamma_c)p(e^{ij}|z_{i\to j}, z_{j\to i}, B) \tag{5.3.11}$$

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 (5.3.11) 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,

$$\begin{aligned} &p(z_{i\to j} = c|D_i, E_i, \boldsymbol{\gamma}, \alpha, \eta, \Psi, \Phi, z_{j\to i} = h) \\ &\propto (n_{i,c} + l_{i\star,}^{-(i\to j)} + \eta\gamma_c)p(e_{ij}|z_{i\to j}, z_{j\to i=h}, B) \\ &= \begin{cases} (n_{i,c} + l_{i\star,c}^{-(i\to j)} + \eta\gamma_c)B_{ch}^{e_{ij}}(1-B_{ch})^{(1-e_{ij})} & \text{for } 1 \leq c \leq C, \\ \eta\gamma_e\frac{\Gamma(e_{ij}+a)\Gamma(1-e_{ij}+b)}{(a+b)\Gamma(a)\Gamma(b)} & \text{for } c = C+1. \end{cases} \end{aligned} \tag{5.3.12}$$

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 content can be written as,

$$L(\{\boldsymbol{\phi}_c\}_{c=1}^{C}, \boldsymbol{\phi}_s) = \sum_{i=1}^{N}\sum_{d=1}^{D} \log P(y_d^i|\boldsymbol{x}_d^i, \boldsymbol{\phi}_{z_d^i}, \boldsymbol{\phi}_s) + \sum_{c=1}^{C} \log p(\boldsymbol{\phi}_c|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) + \log p(\boldsymbol{\phi}_s|\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s^2). \quad (5.3.13)$$

Using a gradient-based optimizer, Eq (5.3.13) 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:

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

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

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 (5.3.8) 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 (5.3.1). 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,

$$\begin{aligned} P(e_{ij} = 1|B, \gamma, \eta, a, b) = \frac{1}{T}\sum_{t=1}^{T}\sum_{g,h} \int d\pi_i \int d\pi_j \int dB_{gh} P(e_{ij} = 1|B_{gh}) \\ p(z_{i\to j} = g|\pi_i)p(z_{j\to i} = h|\pi_j)p(\pi_i|\gamma, \eta)p(\pi_j|\gamma, \eta) \end{aligned} \tag{5.3.14}$$

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.

5.3.4 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 (5.3.8), updating the membership of all the documents imposes a complexity of $\mathcal{O}\big(N\bar{D}(C+M)\big)$, 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 $\mathcal{O}(N^2C)$, 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 $\mathcal{O}(N^2C)$ to $\mathcal{O}(\rho N^2C)$ as ρ usually takes a small value indicating the proportion of deliberate choices. The overall complexity for the proposed algorithm is thus $\mathcal{O}\big(N\bar{D}(C+M)+\rho N^2C\big)$.

• **Summarization.** The proposed HUB model achieves the joint modeling of opinionated content and social network structure. As explicitly expressed in Eq (5.3.8), 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 (5.3.12). 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 $\eta\gamma_c$; 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

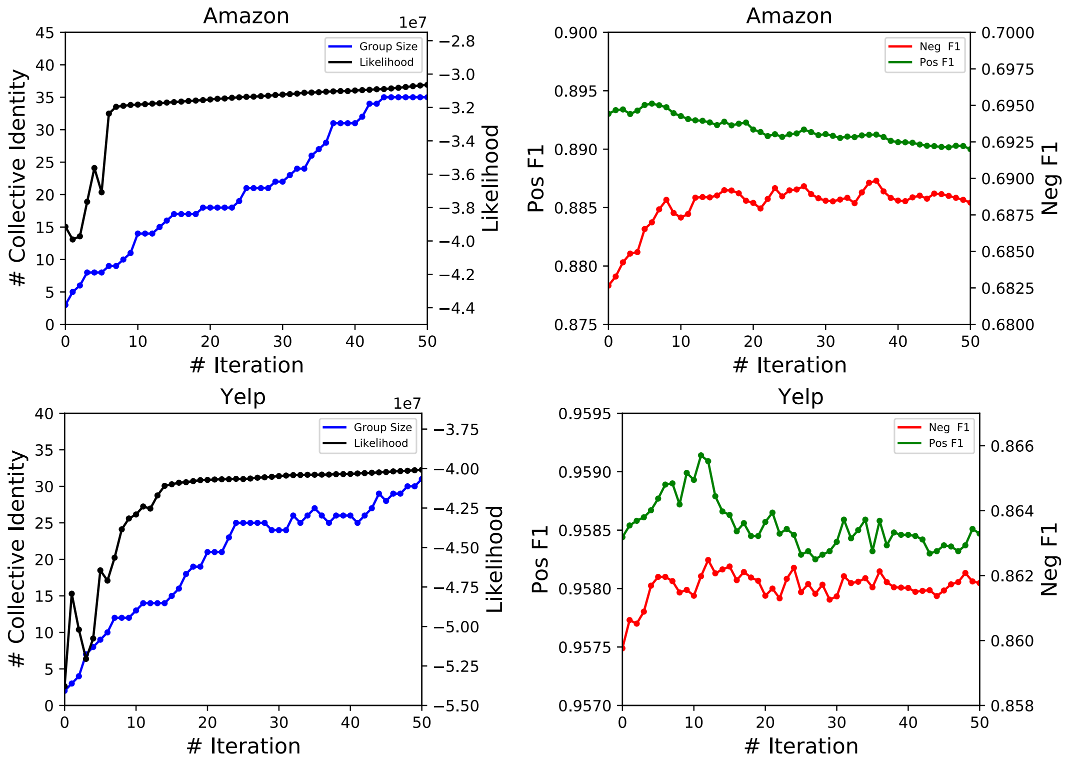

Figure 5.2: Trace of likelihood, model size and sentiment classification performance when training HUB on Amazon and Yelp.

identities, provides pair-wise user affinity that is useful for many personalized applications.

5.4 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.

5.4.1 Datasets

We used two publicly available review datasets collected from Amazon [107] and Yelp [1], for our evaluation purpose. We directly utilized the processed text data as introduced in Section 4.5. As for the network structure, Yelp dataset provides user friendship imported from users' Facebook friend connections, while there is no explicit social network in Amazon dataset. We utilized the

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

"co-purchasing" information to build the network structure for Amazon users. We have 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. Both the source codes and data are available online [2].

5.4.2 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 5.2. 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 content, 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 5.3.

[2] JNET. https://github.com/Linda-sunshine/HUB.

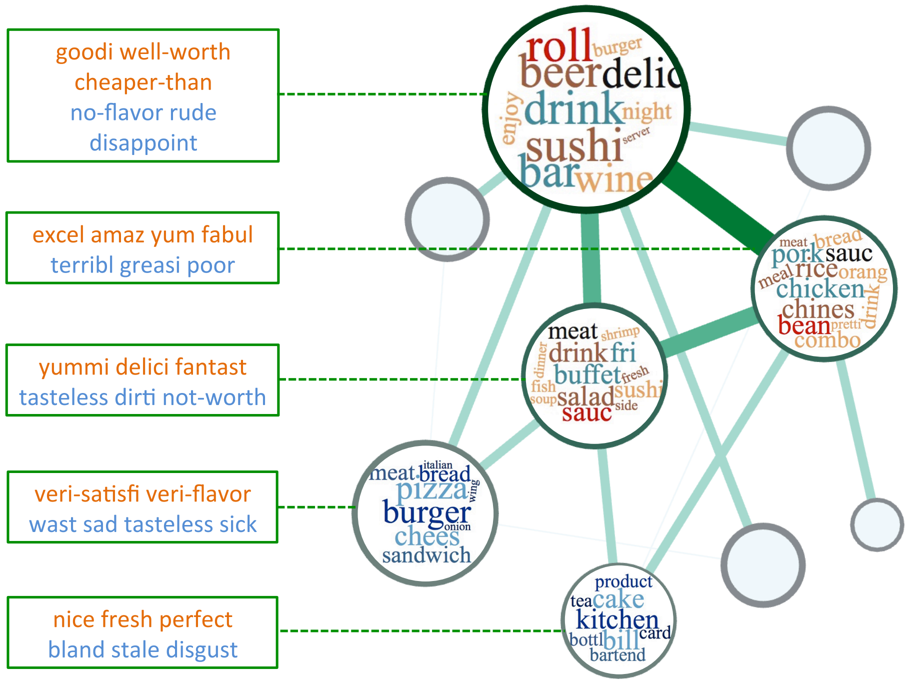

Figure 5.3: The identified behavior patterns among a subset of collective identities on Yelp dataset.

In Figure 5.3, 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.

5.4.3 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 [39], 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 [24] 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 [41] 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 content and the sentiment labels, while social connections are neglected. 5) **Graph-Based Semi-Supervised Learning (GBSSL)** : we followed [132] 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 content 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 6.3. 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 each user's mixed membership simply based on review content, 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 content 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

Table 5.1: 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.

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.

5.4.4 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,

$$sim(u_i, u_j) = \sum_{g=1}^{G} \sum_{h=1}^{H} u_{i,g} \cdot u_{j,h} \cdot B_{gh} \tag{5.4.1}$$

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++ [133] and factorization machine (FM) [134] 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 5.2. 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 content. Due to the very sparse distribution of items in both datasets, matrix factorization based methods suffer in performance.

Table 5.2: 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**

5.4.5 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 content 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 content to infer their personal identity, thus to provide helpful friend recommendation.

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 content. 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 (5.3.14).

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 5.3 by comparing against a random solution, i.e., divide the performance by the performance of a random recommendation.

Table 5.3: 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**

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.

5.5 Conclusion

In the work, 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.

Though text and network are jointly considered, they are only correlated by sharing the same mixing component, without explicitly modeling the mutual influence between them. With such explicit modeling of correlations between text and network enabled, we can get a better understanding of users' diverse behaviors.

Chapter 6

User Representation Learning with Joint Network Embedding and Topic Embedding

In this Chapter, we propose to perform user representation learning via ***explicit modeling of the structural dependency*** among different modalities of user-generated data, thus to better understand user intents. In particular, we developed a probabilistic generative model to learn user embeddings via a joint modeling of text content and network structure. In order to model user-generated text content, we further embed topics to the same latent space as users to enable a joint network embedding and topic embedding and capture the relationships explicitly. We evaluated the proposed solution on a large collection of Yelp reviews and StackOverflow discussion posts, with the associated network structures. The proposed model outperformed several state-of-the-art topic modeling based user models with better predictive power in unseen documents, and state-of-the-art network embedding based user models with better link prediction in unseen nodes. The learned user representations are also proved to be useful in content recommendation, e.g., expert finding in StackOverflow.

6.1 Introduction

User modeling builds up conceptual representations of users, which helps automated systems to better capture users' needs and to create compelling experience for them [135, 136]. Thanks to the rapid development of social media, ordinary users can actively participate in online activities and create vast amount of observational data, such as social interactions [56, 137] and opinionated text content [138–140], which in turns provides informative clues about their intents. Extensive research efforts have proved the value of user representation learning in various real-world applications, such as latent factor models for collaborative filtering [134, 141], topic models for content modeling [50, 142], embedding based models for social link prediction [22, 65], etc.

However, most work focuses on one particular type of behavior signals for user modeling, where the dependency across multiple types of behavior data is ignored. For example, users' social interactions [16, 65] and their generated text data [48, 50, 142] have been extensively studied, but they are mostly modeled in isolation. Even among a few attempts for joint modeling of different types of user-generated data [79, 138], ***the explicit modeling of dependency*** among multiple behavior modalities is still missing. For example, Yang et al. [79] incorporated user-generated text content into network representation learning via a joint matrix factorization. In their solution, text data modeling is only used as a regularization for network modeling; and thus the learnt model is not in a position to predict future text content. Gong and Wang [138] paired the task of textual sentiment classification with that of social network modeling, and represented each user as a mixture over the instances of these paired tasks. Though text and network are jointly considered, they are only correlated by sharing the same mixing component, without explicitly modeling the mutual influence between them.

In social psychology and cognitive science, ***schema*** defines the knowledge structure a person holds that organizes categories of information and the relationships among them [143]. In other words, schema provides a way to transform input information to a generalized representation that realizes the dependency between different categories of information. Inspired from this concept, we propose to construct a uniform low-dimensional space to capture user scheme, which preserves the properties of each modality of user-generated data, so as to capture the dependency among them. The space should be constructed in such a way that the closeness among different modalities of user-generated data can be easily characterized by the similarity measured in the latent space. For example, connected users in a social network should be closer to each other in this latent space; and by mapping user behavior data into this space, e.g., text data, users should be surrounded by their own generated data.

To realize this new perspective of user representation learning, we exploit two most widely available and representative forms of user-generated data, i.e., text content and social interactions. We develop a probabilistic generative model to integrate user modeling with content and network embedding. Due to the unstructured nature of text, we appeal to topic models to represent user-generated text content [48, 50]. We embed both users and topics to the same low-dimensional space to capture their mutual dependency. On one hand, a user's affinity to a topic is characterized by his/her proximity to the topic in this latent space, which is utilized to generate each text document of the user. On the other hand, the affinity between users is directly modeled by the proximity between users' embedding vectors, which is utilized to generate the corresponding social network connections. In this latent space, the two modalities of user-generated data are correlated explicitly. The user representation is obtained by posterior inference over a set of training data, via variational Bayesian. To reflect the nature of our proposed user representation learning solution, we name the solution **J**oint **N**etwork **E**mbedding and **T**opic Embedding, or JNET for short.

Extensive experiments are performed on two large collections of user-generated text documents from Yelp and StackOverflow, together with their network structures. Compared with a set of state-of-the-art user representation learning solutions, from the perspective of content modeling [48, 142, 144] or network modeling [16, 79], clear advantages of JNET are observed by its explicit modeling of correlations among different categories of information enabled by the latent space. The use of learnt user representation generalizes beyond content prediction and link prediction: it accurately suggests technical discussion threads for users to participate in StackOverflow, e.g., expert recommendation.

6.2 Joint Network Embedding and Topic Embedding

6.2.1 Model Specification

We focus on user representation learning based on user-generated text data, in conjunction with their social network interactions. Formally, denote a collection of U users as $\mathcal{U} = \{u_1, u_2, ...u_U\}$, in which each user u_i is associated with a set of D_i text documents $\mathcal{D}_i = \{\boldsymbol{x}_{i,d}\}_{d=1}^{D_i}$. Each document $\boldsymbol{x}_d$ is represented as a bag of words $\boldsymbol{x}_d = \{w_1, w_2, .., w_N\}$, where w_n is chosen from a vocabulary of fixed size V. Each user is also associated with a set of social connections, referred as friendship, which we denote as $\mathcal{E}_i = \{e_{ij} = 1\}_{j \neq i}^{U}$. For a pair of users u_i and u_j, the binary observation e_{ij} denotes the connection between them: $e_{ij} = 1$ indicates they are directly connected in the network, i.e., friends; otherwise, $e_{ij} = 0$.

We represent each user as a real-valued continuous vector $\boldsymbol{u}_i \in \mathbb{R}^M$ in a low-dimensional space. And we seek to impose a joint distribution over the observations in each user's associated text documents and social interactions, so as to capture the underlying structural dependency between these two types of data. Based on our assumption that both types of users-generated data are governed by the same underlying user intent, we explicitly model the joint distribution as $p(\mathcal{D}_i, \mathcal{E}_i) = \int p(\mathcal{D}_i, \mathcal{E}_i, \boldsymbol{u}_i) \mathrm{d}\boldsymbol{u}_i$, which can be further decomposed into $p(\mathcal{D}_i, \mathcal{E}_i, \boldsymbol{u}_i) = p(\mathcal{D}_i | \mathcal{E}_i, \boldsymbol{u}_i) p(\mathcal{E}_i | \boldsymbol{u}_i) p(\boldsymbol{u}_i)$. We assume given the user representation $\boldsymbol{u}_i$, the generation of text documents in $\mathcal{D}_i$ is independent from the generation of social interactions in $\mathcal{E}_i$, i.e., $p(\mathcal{D}_i | \mathcal{E}_i, \boldsymbol{u}_i) = p(\mathcal{D}_i | \boldsymbol{u}_i)$. As a result, the modeling of joint probability over a user's observational data with his/her latent representation can be decomposed into three related modeling tasks: 1) $p(\mathcal{D}_i | \boldsymbol{u}_i)$ for content modeling, 2) $p(\mathcal{E}_i | \boldsymbol{u}_i)$ for social connection modeling, and 3) $p(\boldsymbol{u}_i)$ for user embedding modeling.

We appeal to statistical topic models [47, 48] for content modeling, because of their impressive effectiveness shown in existing empirical studies. Classical topic models view each topic as a discrete set of indices, from which specific word distributions are sampled from. But it is not compatible with our continuous user representation. To build direct connection between users and topics, we decide to embed topics into the same latent space as users. By projecting topic embedding vectors to each user's embedding vector, we can easily measure each user's affinity to each topic, thus to capture users' topical preferences. Another benefit is that it also allows us to measure the topical variance in documents from the same user and establish a valid predictive distribution of his/her documents.

Formally, we assume there are in total K topics underlying the corpus with each one represented as an embedding vector $\boldsymbol{\phi}_k \in \mathbb{R}^M$ in the same latent space; denote $\Phi \in \mathbb{R}^{K \times M}$ to facilitate our representation of each user's affinity towards different topics, i.e., $\Phi \cdot \boldsymbol{u}_i$. In order to be qualified as a distribution over topics, the vector $\Phi \cdot \boldsymbol{u}_i$ has to lie in a K-dimensional simplex. Thus, we use a logistic-normal distribution to map $\Phi \cdot \boldsymbol{u}_i$ back to the simplex [145]. As this mapped vector reflects user u_i's topical preferences, it serves as the prior of topic distribution in each

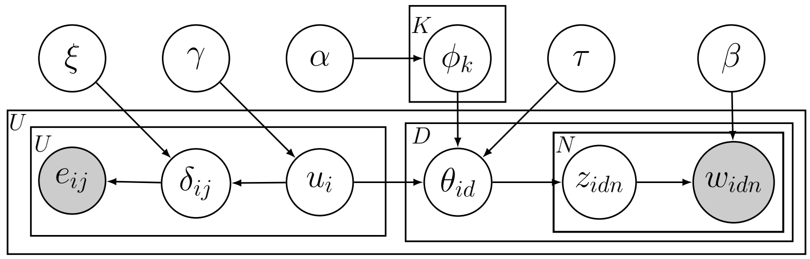

Figure 6.1: Graphical model representation of JNET. The upper plate indexed by K denotes the learnt topic embeddings. The outer plate indexed by U denotes distinct users in the collection. The inner plates indexed by U and D denote each user's social connections and text documents respectively. The inner plate indexed by N denotes the word content in one text document.

text document from him/her. Specifically, denote the document-level topic vector as $\boldsymbol{\theta}_{id} \in \mathbb{R}^K$, we have $\boldsymbol{\theta}_{id} \sim \mathcal{N}(\Phi \cdot \boldsymbol{u}_i, \tau^{-1}I)$, where τ characterizes the uncertainty when user u_i is choosing topics from his/her global topic preferences for each single document. By projecting the document-level topic vector into the probability simplex, we obtain the topic distribution for document $\boldsymbol{x}_{i,d}$: $\boldsymbol{\pi}_{id} = \text{softmax}(\boldsymbol{\theta}_{id})$, from which we sample a topic indicator $z_{idn} \in \{1, ..., K\}$ for each word w_{idn} in $\boldsymbol{x}_{i,d}$ by $z_{idn} \sim \text{Multi}(\text{softmax}(\boldsymbol{\theta}_{id}))$. As in conventional topic models, each topic k is also associated with a multinomial distribution $\boldsymbol{\beta}_k$ over a fixed vocabulary, and each word w_{idn} is then drawn from respective word distribution indicated by corresponding topic assignment, i.e., $w_{idn} \sim p(w|\boldsymbol{\beta}_{z_{idn}})$. Put all pieces together, the task of content modeling for each user can be summarized as $p(\mathcal{D}_i|\boldsymbol{u}_i) = \prod_{d=1}^{D_i} p(\boldsymbol{\theta}_{id}|\boldsymbol{u}_i, \Phi, \tau) \prod_{n=1}^{N} p(z_{idn}|\boldsymbol{\theta}_{id}) p(w_{idn}|z_{idn}, \boldsymbol{\beta})$.

The key in modeling social connections is to understand the closeness among users. As we represent users with a real-valued continuous vector, this can be easily measured by the vector inner product in the learnt low-dimensional space. Define the underlying affinity between a pair of users u_i and u_j as δ_{ij}, we assume $\mathbb{E}[\delta_{ij}] = \boldsymbol{u}_i^\top \boldsymbol{u}_j$. To capture uncertainty of the affinity between different pairs of users, we further assume δ_{ij} is drawn from a Gaussian distribution centered at the measured closeness, $\delta_{ij} \sim \mathcal{N}(\boldsymbol{u}_i^\top \boldsymbol{u}_j, \xi^2)$, where ξ characterizes the concentration of this distribution. The observed social connection e_{ij} between user u_i and u_j is then assumed as a realization of this underlying user affinity: $e_{ij} \sim \text{Bernoulli}(\text{logistic}(\delta_{ij}))$. As a result, the task of social connection modeling can be achieved by $p(\mathcal{E}_i|\boldsymbol{u}_i) = \prod_{j \neq i}^{U} p(e_{ij}|\delta_{ij}) p(\delta_{ij}|\boldsymbol{u}_i, \boldsymbol{u}_j)$.

We do not have any specific constraint on the form or distribution of latent user embedding vectors $\{\boldsymbol{u}_i\}_{i=1}^{U}$, as long as they are in a M-dimensional space. For simplicity, we assume they are drawn from an isotropic Gaussian distribution, i.e., $\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, \gamma^{-1}\boldsymbol{I})$, where γ measures the concentration of different users' embedding vectors. Other types of prior distribution can also be introduced, if one has more knowledge about the user embeddings, e.g., sparsity or a particular geometric shape.

But it is generally preferred to have conjugate priors, so as to simplify later posterior inference steps. This concludes our user embedding modeling task.

To make our model a complete generative model, we also impose flat priors over the topic embedding, i.e., , $\boldsymbol{\phi}_k \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\boldsymbol{I})$. Putting these components together, the generative process of our solution can be described as follows:

- For each topic $\boldsymbol{\phi}_k$:
  - Draw a topic compact representation $\boldsymbol{\phi}_k \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\boldsymbol{I})$
- For each user u_i:
  - Draw user compact representation $\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, \gamma^{-1}\boldsymbol{I})$
  - For every other user u_j:
    * Draw affinity δ_{ij} between $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$, $\delta_{ij} \sim \mathcal{N}(\boldsymbol{u}_i^{\top}\boldsymbol{u}_j, \xi^2)$
    * Draw interaction e_{ij} between $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$, $e_{ij} \sim$ Bernoulli(logistic(δ_{ij}))
- For each document of user u_i:
  - Draw the user-document topic preference vector $\boldsymbol{\theta}_{id} \sim \mathcal{N}(u_i \cdot \Phi, \tau^{-1}\boldsymbol{I})$
  - For each word w_{idn}:
    * Draw topic assignment $z_{idn} \sim$ Multi(softmax($\boldsymbol{\theta}_{id}$))
    * Draw word $w_{idn} \sim$ Multi($\boldsymbol{\beta}_{z_{idn}}$)

In this generative process, we make two explicit assumptions:

- the dimensionality M of the compact representation of topics and users is predefined and fixed;
- the word distributions under topics are parameterized by a $K \times V$ matrix $\boldsymbol{\beta}$ where $\boldsymbol{\beta}_{kv} = p(w^v = 1|z^k = 1)$ over a fixed vocabular of size V.

We obtain a full generative model capturing the structural dependency among multiple modalities of user-generated data for effective user representation learning. In essence, we are performing a **J**oint **N**etwork **E**mbedding and **T**opic Embedding, we name the resulting model as JNET in short; we illustrate our imposed dependency between different components of JNET in Figure 6.1, using a graphical model representation.

6.2.2 Variational Bayesian Inference

The compact user representations can be obtained via posterior inference over the latent variables u on a given set of data. However, posterior inference is not analytically tractable in JNET due to the coupling among several latent variables, i.e., user-user affinity δ, user embedding u, topic embedding Φ, document-level topic proportion $\boldsymbol{\theta}$ and topic assignment of each word z. We appeal to a mean-field variational method to efficiently approximate the posterior distributions. Different from the conjugate-exponential family pairs, non-conjugate logistic-normal priors lead to difficulty in lower bound based approximation. We further extend the approximation using Taylor expansion [145] to improve the quality of inferred posteriors.

We begin by postulating a factorized distribution:
$q(\Phi, U, \Delta, \Theta, Z) = \prod_{k=1}^{K} q(\boldsymbol{\phi}_k) \prod_{i=1}^{U} q(\boldsymbol{u}_i) \Big[\prod_{j=1, j\neq i}^{U} q(\delta_{ij}) \prod_{d=1}^{D} q(\boldsymbol{\theta}_{id}) \prod_{n=1}^{N} q(z_{idn}) \Big]$,
where the factors have the following parametric forms:

$$
\begin{aligned}
&q(\boldsymbol{\phi}_k) = \mathcal{N}(\boldsymbol{\phi}_k | \boldsymbol{\mu}^{(\phi_k)}, \Sigma^{(\phi_k)}), q(\boldsymbol{u}_i) = \mathcal{N}(\boldsymbol{u}_i | \boldsymbol{\mu}^{(u_i)}, \Sigma^{(u_i)}),\\
&q(\delta_{ij}) = \mathcal{N}(\delta_{ij} | \mu^{(\delta_{ij})}, {\sigma^{(\delta_{ij})}}^2), q(\boldsymbol{\theta}_{id}) = \mathcal{N}(\boldsymbol{\theta}_{id} | \boldsymbol{\mu}^{(\theta_{id})}, \Sigma^{(\theta_{id})}),\\
&q(z_{idn}) = \text{Mult}(z_{idn} | \eta_{idn})
\end{aligned}
$$

Because the topic proportion vector $\boldsymbol{\theta}_{id}$ is inferred in each document, it is not necessary to estimate a full covariance matrix for it [145]. Hence, in its variational distribution, we only estimate the diagonal variance parameters.

Variational algorithms aim to minimize the KL divergence from the approximated posterior distribution q to the true posterior distribution p. It is equivalent to tightening the evidence lower bound

(ELBO) by Jensen's inequality [48]:

$$
\begin{aligned}
&\log p(\boldsymbol{w}, \boldsymbol{e}, \Phi, U, \Delta, \Theta, Z|\alpha, \beta, \gamma, \tau) \\
&\geq \mathbb{E}_q[\log p(U, \Theta, Z, \Phi, \Delta, \boldsymbol{w}, \boldsymbol{e}|\alpha, \beta, \gamma, \tau)] - \mathbb{E}_q[\log q(U, \Theta, Z, \Phi, \Delta)]
\end{aligned} \tag{6.2.1}
$$

where the expectation is taken with respect to the factorized variaitonal distribution of latent variables $q(\Phi, U, \Delta, \Theta, Z)$.

Let $\mathcal{L}(q)$ denote the right-hand side of Eq (6.2.1), the first step to maximize this lower bound is to derive the analytic form of posterior expectations required in $\mathcal{L}(q)$. Thanks to the conjugate priors introduced on $\{\boldsymbol{u}_i\}_{i=1}^U, \Phi = \{\boldsymbol{\phi}_k\}_{k=1}^K$, the expectations related to these latent variables have closed form solutions. Since there is no conjugate prior for logistic Normal distribution, we use Taylor expansions to approximate the expectations related to $\boldsymbol{\theta}_{id}, \delta_{ij}$. Next we describe the detailed inference procedure for each latent variable, but due to space limit we will omit most of details for the expectation calculation.

• **Estimate topic embedding.** For each topic k, we relate the terms associated with $q(\boldsymbol{\phi}_k|\boldsymbol{\mu}^{(\phi_k)}, \Sigma^{(\phi_k)})$ in Eq (6.2.1) and take maximization w.r.t. $\boldsymbol{\mu}^{(\phi_k)}$ and $\Sigma^{(\phi_k)}$. Closed form estimations of $\boldsymbol{\mu}^{(\phi_k)}, \Sigma^{(\phi_k)}$ exist,

$$
\boldsymbol{\mu}^{(\phi_k)} = \tau \Sigma^{(\phi_k)} \sum_{i=1}^{U} \sum_{d=1}^{D_i} \boldsymbol{\mu}_k^{(\theta_{id})} \boldsymbol{\mu}^{(u_i)} \tag{6.2.2}
$$

$$
\Sigma^{(\phi_k)} = \left[\alpha \boldsymbol{I} + \tau \sum_{i=1}^{U} \sum_{d=1}^{D_i} (\Sigma^{(u_i)} + \boldsymbol{\mu}^{(u_i)} \boldsymbol{\mu}^{(u_i)^{\mathsf{T}}})\right]^{-1} \tag{6.2.3}
$$

As we can find from Eq (6.2.3), the estimation of $\Sigma^{(\phi_k)}$ is not related to a specific topic k, because we impose an isotropic Gaussian prior for all $\{\boldsymbol{\phi}_k\}_{k=1}^K$ in JNET. It suggests that the correlations between different topic embedding dimensions are homogeneous across topics. Thus, $\{\boldsymbol{\mu}^{(\phi_k)}\}_{k=1}^K$ serve as a posterior mode estimation of topic embeddings for the current corpora. Interestingly, as we can notice that the posterior covariance $\Sigma^{(\phi_k)}$ of topic embeddings is only related to user embeddings, which introduces information from network modeling into content modeling.

• **Estimate user embedding.** For each user i, we relate the terms associated with $q(\boldsymbol{u}_i|\boldsymbol{\mu}^{(u_i)}, \Sigma^{(u_i)})$ in Eq (6.2.1) and maximize it with respect to $\boldsymbol{\mu}^{(u_i)}, \Sigma^{(u_i)}$. The procedure here is similar to the aforementioned posterior topic embedding estimation. Closed form estimations can also be achieved for these two parameters as follows:

$$\boldsymbol{\mu}^{u_i} = \Sigma^{(u_i)}\big(\tau \sum_{d=1}^{D_i}\sum_{k=1}^{K} \boldsymbol{\mu}_k^{(\theta_{id})}\boldsymbol{\mu}^{(\phi_k)} + \sum_{j\neq i}^{U} \frac{\mu^{(\delta_{ij})}}{\xi^2}\mu^{(u_j)}\big) \tag{6.2.4}$$

$$\Sigma^{(u_i)} = \Big[\gamma \boldsymbol{I} + \tau D_i \sum_{k=1}^{K}(\Sigma^{(\phi_k)} + \boldsymbol{\mu}^{(\phi_k)}\boldsymbol{\mu}^{(\phi_k)^{\mathsf{T}}}) + \sum_{j\neq i}^{U}\frac{1}{\xi^2}(\Sigma^{(u_j)} + \boldsymbol{\mu}^{(u_j)}\boldsymbol{\mu}^{(u_j)^{\mathsf{T}}})\Big] \tag{6.2.5}$$

The effect of joint content modeling and network modeling for user representation learning is clearly depicted in this posterior estimation of user embedding vectors. The updates of $\boldsymbol{\mu}^{(u_i)}$ and $\Sigma^{(u_i)}$ come from two types of influence: the textual content and social interactions of the current user. For example, the posterior mode estimation of user embedding vector $\boldsymbol{u}_i$ is a weighted average over the topic vectors that this user has used in his/her past text documents and the user vectors from his/her friends. And the weights measure his/her affinity to those topics and users in each specific observation. The updates exactly reflect the formation of "User Schema" in social psychology, which can be understood from two perspectives: both modalities of user-generated data contribute to the shaping of user embedding, while the structural dependency between them are reflected in this unified user representation.

• **Estimate per-document topic proportion vector.** Similar procedures as above can be taken to estimate $\boldsymbol{\mu}^{(\theta_{id})}$ and $\Sigma^{(\theta_{id})}$. Due to the lack of conjugate prior for logistic Normal distributions, we apply Taylor expansion and introduce an additional free variational parameter ζ in each document. Because there is no closed form solution for the resulting optimization problem, we use gradient ascent to optimize $\boldsymbol{\mu}^{(\theta_{id})}$ and $\Sigma^{(\theta_{id})}$ with the following gradients,

$$\frac{\partial L}{\partial \boldsymbol{\mu}_k^{(\theta_{id})}} = -\tau\boldsymbol{\mu}_k^{(\theta_{id})} + \tau \sum_{m=1}^{M} \boldsymbol{\mu}_m^{(\phi_k)}\boldsymbol{\mu}_m^{(u_i)} + \sum_{n=1}^{N}\big[\eta_{idnk} - \zeta^{-1}\exp(\boldsymbol{\mu}_k^{(\theta_{id})} + \frac{1}{2}\Sigma_{kk}^{(\theta_{id})})\big] \tag{6.2.6}$$

$$\frac{\partial L}{\partial \Sigma_{kk}^{(\theta_{id})}} = -\tau - \frac{N}{\zeta}\exp(\boldsymbol{\mu}_k^{(\theta_{id})} + \frac{1}{2}\Sigma_{kk}^{(\theta_{id})}) + \frac{1}{\Sigma_{kk}^{(\theta_{id})}} \tag{6.2.7}$$

where $\zeta = \sum_{k=1}^{K}\exp(\mu_k^{(\boldsymbol{\theta}_{id})} + \frac{1}{2}\Sigma_{kk}^{(\theta_{id})})$. As we mentioned before, only the diagonal elements in $\Sigma_{id}^{(\theta)}$ are statistically meaningful (i.e., variance). And we simply set its off-diagonal elements to zero in gradient update. As the variance has to be non-negative, we can instead estimate the square root of it to avoid solving a constrained optimization problem. The gradient function suggests that the document-level topic proportion vector should align with the corresponding compact user representation and topic representation. Although no closed form estimations of $\boldsymbol{\mu}^{(\theta_{id})}$ and $\Sigma^{(\theta_{id})}$ exist, the expected property of $\boldsymbol{\mu}^{(\theta_{id})}$ is clearly reflected in Eq (6.2.6): the proportion of each topic in document $\boldsymbol{x}_{i,d}$ should align with this user's preference to this topic (i.e., affinity in the embedding space), and the topic assignment in document content. And the variance is introduced by the uncertainty of per-word topic choice and the intrinsic uncertainty of a user's affinity with a topic.

• **Estimate user affinity.** Similar approach can be applied here to estimate $\mu^{(\delta_{ij})}$ and ${\sigma^{(\delta_{ij})}}^2$ which govern the latent user affinity. Again, gradient ascent is utilized to optimize $\mu^{(\delta_{ij})}$ and $\Sigma^{(\theta_{id})}$,

$$\frac{\partial L}{\partial \mu^{(\delta_{ij})}} = e_{ij} - \frac{1}{\varepsilon}\exp\left(\mu^{(\delta_{ij})} + \frac{1}{2}{\sigma^{(\delta_{ij})}}^2\right) - \frac{\mu^{(\delta_{ij})}}{\xi^2} + \frac{{\boldsymbol{\mu}_m^{(u_i)}}^{\mathsf{T}}\boldsymbol{\mu}_m^{(u_j)}}{\xi^2} \tag{6.2.8}$$

$$\frac{\partial L}{\partial \sigma^{(\delta_{ij})}} = -\frac{\sigma^{(\delta_{ij})}}{\varepsilon}\exp\left(\mu^{(\delta_{ij})} + \frac{1}{2}{\sigma^{(\delta_{ij})}}^2\right) - \frac{\sigma^{(\delta_{ij})}}{\xi^2} + \frac{1}{\sigma^{(\delta_{ij})}} \tag{6.2.9}$$

The gradient functions clearly indicate the latent affinity between a pair of users is closely related with their observed connectivity and their closeness in the embedding space.

• **Estimate word topic assignment.** The topic assignment z_{idn} for each word w_{idn} in each document $\boldsymbol{x}_{i,d}$ can be easily estimated by,

$$\eta_{idnk} \propto \exp\{\mu_k^{(\boldsymbol{\theta}_{id})} + \sum_{v=1}^{V} w_{idnv}\log\boldsymbol{\beta}_{kv}\} \tag{6.2.10}$$

We execute the above variational inference procedures in an alternative fashion until the lower bound $\mathcal{L}(q)$ defined in Eq (6.2.1) converges. The variational inference algorithm endows rich independence structures between the variational parameters, allowing straightforward parallel computing. Since the variational parameters can be grouped into document level: $\boldsymbol{\mu}^{(\theta_{id})}$, $\Sigma^{(\theta_{id})}$ and η, topic-level: $\boldsymbol{\mu}^{(\phi_k)}$ and $\Sigma^{(\phi_k)}$, and user-level: $\boldsymbol{\mu}^{(u_i)}$, $\Sigma^{(u_i)}$, $\mu^{(\delta_{ij})}$ and ${\sigma^{(\delta_{ij})}}^2$, we perform alternative update in parallel to improve computational efficiency, e.g., fix topic-level parameters and user-level parameters, and distribute the documents across different CPUs to estimate their own $\boldsymbol{\mu}^{(\theta_{id})}$, $\Sigma^{(\theta_{id})}$ and η in parallel for large collections of user-generated data.

6.2.3 Parameter Estimation

When performing the variational inference described above, we have assumed the knowledge of model parameters $\alpha, \gamma, \tau, \xi$ and β. Based on the inferred posterior distribution of latent variables in

JNET, these model parameters can be readily estimated by the Expectation-Maximization (EM) algorithm.

Among these model parameters, the most important ones are the priors for user embedding γ and topic embedding α, and word-topic distribution $\boldsymbol{\beta}$. As ξ and τ serve as the variance for user affinity δ_{ij} and document topic proportion vector $\boldsymbol{\theta}_{id}$, and we have large amount of relevant observations in users' documents and social connections, our model is less sensitive to their settings. Therefore, we will estimate α, γ and β with respect to available training data, and empirically set ξ and τ.

By taking the gradient of $\mathcal{L}(q)$ in Eq (6.2.1) with respect to α, and set it to 0, we get the closed form estimation of α as follows:

$$\alpha = \frac{KM}{\sum_{k=1}^{K}[\sum_{m=1}^{M}\Sigma_{mm}^{(\phi_k)} + \boldsymbol{\mu}^{(\phi_k)}{}^{\mathsf{T}}\boldsymbol{\mu}^{(\phi_k)}]}, \tag{6.2.11}$$

Similarly, the closed form estimation of γ can be easily derived as,

$$\gamma = \frac{UM}{\sum_{i=1}^{U}[\sum_{m=1}^{M}\Sigma_{mm}^{(u_i)} + \boldsymbol{\mu}^{(u_i)}{}^{\mathsf{T}}\boldsymbol{\mu}^{(u_i)}]}. \tag{6.2.12}$$

And the closed form estimation for word-topic distribution β can be achieved by,

$$\boldsymbol{\beta}_{kv} \propto \sum_{i=1}^{U}\sum_{d=1}^{D_i}\sum_{n=1}^{N} w_{idnv}\boldsymbol{\eta}_{idnv}, \tag{6.2.13}$$

where w_{idnv} indicates the nth word in u_i's dth document is the vth word in the vocabulary.

The resulting EM algorithm consists of E-step and M-step. In E-step, variational parameters are inferred based the procedures described in Section 6.2.2 until convergence; and in M-step, model parameters α, γ and β are estimated based on collected sufficient statistics from E-step. These two steps are repeated until the lower bound $\mathcal{L}(q)$ converges over all training data.

Inferring the latent variables with each user and each topic are computationally cheap. Specifically, by Eq (6.2.3), updating the variables associated with each topic imposes a complexity of $\mathcal{O}\big(KM^2|D|\big)$, where K is the total numnber of topicss, M is the latent dimension, $|D|$ is the total number of documents. By Eq (6.2.4), updating the variables for each user imposes a complexity of $\mathcal{O}(M^2U^2)$ where U is the total number of users. Estimating the latent variables for per-document topic proportion imposes a complexity of $\mathcal{O}(|D|K(\bar{N}+M))$ by Eq (6.2.6), where $\bar{N}$ is the average document length.

And updating variables for each pair of user affinity takes constant time while there are U^2 affinity variables. With the consideration of the total number of users and topics, the overall complexity for the proposed algorithm is $\mathcal{O}(KM^2|D| + M^2U^2)$ by ignoring constants.

6.3 Experiments

We evaluated the proposed model on large collections of Yelp reviews and StackOverflow forum discussions, together with their user network structures. Qualitative analysis demonstrates the descriptive power of JNET through direct mapping of user vectors and topic vectors into 2-D space. A set of quantitative evaluations confirm the effectiveness of JNET compared with several existing state-of-the-art user representation learning solutions: social connections enhance the model's predictive power in unseen documents, and text content facilitates the model's prediction on missing links. Moreover, the learnt user representation enables accurate content recommendation to users.

6.3.1 Experimental Setup

• **Datesets.** We employed two large publicly available user-generated text datasets together with the associated user networks: 1) **Yelp**, collected from Yelp dataset challenge[1], is a collection of 187,737 Yelp restaurant reviews generated by 10,830 users. The Yelp dataset provides user friendship imported from their Facebook friend connections. Among the whole set of users, 10,194 of them have friends with an average of 10.65 friends per user. 2) **StackOverflow**, collected from Stackoverflow[2], consists of 244,360 forum discussion posts generated by 10,808 users. While there is no explicit network structure in StackOverflow dataset, we utilized the "***reply-to***" information in the discussion threads to build the network, because this relation suggests implicit social connections among users, e.g., their expertise and technical topic interest. We ended up with 10,041 users having friends, with an average of 5.55 connections per user. We constructed 5,000 unigram and bigram text features based on Document Frequency (DF) in both datasets. We randomly split the data for 5-fold cross validation in all the reported experiments.

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

• **Baselines.** We compared the proposed JNET model against a rich set of user representation learning methods: topic modeling based solutions, the state-of-the-art network embedding methods, together with works combining text and network for learning user embeddings.

1) **Latent Dirichlet Allocation (LDA)** [48] assumes a global conjugate Dirichlet prior distribution to generate the topic distribution in documents across different users. This classic topic model does not explicitly model user, thus user embedding is created by averaging the posterior topic proportion of documents associated with a user. 2) **Relational Topic Model (RTM)** [144] explicitly models the connection between two documents conditioned on their text content. In our application scenario, instead of document-level network, we constructed a user-level network by concatenating all documents from one user into one to fit this baseline model. 3) **Hidden Factors as Topics (HFT)** [142] combines latent rating dimensions of users captured by latent-factor based recommendation model, with latent review topics learnt from topic models to achieve better recommendation. Latent user factors indicating users' implicit tastes and latent item factors representing properties of items are jointly learnt from text and rating information. Since StackOverflow dataset does not come with natural ratings, users' "***upvote***" toward a question is utilized as a proxy of rating. 4) **Collaborative Topic Regression (CTR)** [50] combines collaborative filtering with topic modeling to explain the observed text content and ratings, which provides latent structures for both users and items. Since CTR assumes one document for each item, we constructed an item by aggregating one restaurant's reviews in Yelp and one question's answers in StackOverflow respectively. 5) **DeepWalk (DW)** [16] takes network structure as input to learn social representations of vertices in the network. Via techniques from neural language models, DW uses local information obtained from truncated random walks to learn latent node representations by treating walks as an equivalent of sentences. 6) **Text-Associated DeepWalk (TADW)** [79] further incorporates text features of vertices into network representation learning under the framework of joint matrix factorization.

Among these baselines, LDA learns topic representation using documents only; DW learns user representation by network structure only; CTR and HFT consider both user factors and item factors (a.k.a. restaurant in Yelp and question in Stackoverflow) in modeling text content and ratings; TADW and RTM learn user representations by jointly modeling text content and network structure.

• **Parameter Settings.** We set the latent dimensions of user and topic embeddings to 10 in JNET and baselines. Though a larger dimension might improve some specific tasks' performance, it is computationally more expensive. As we tuned the topic size from 10 to 100, we found the learnt

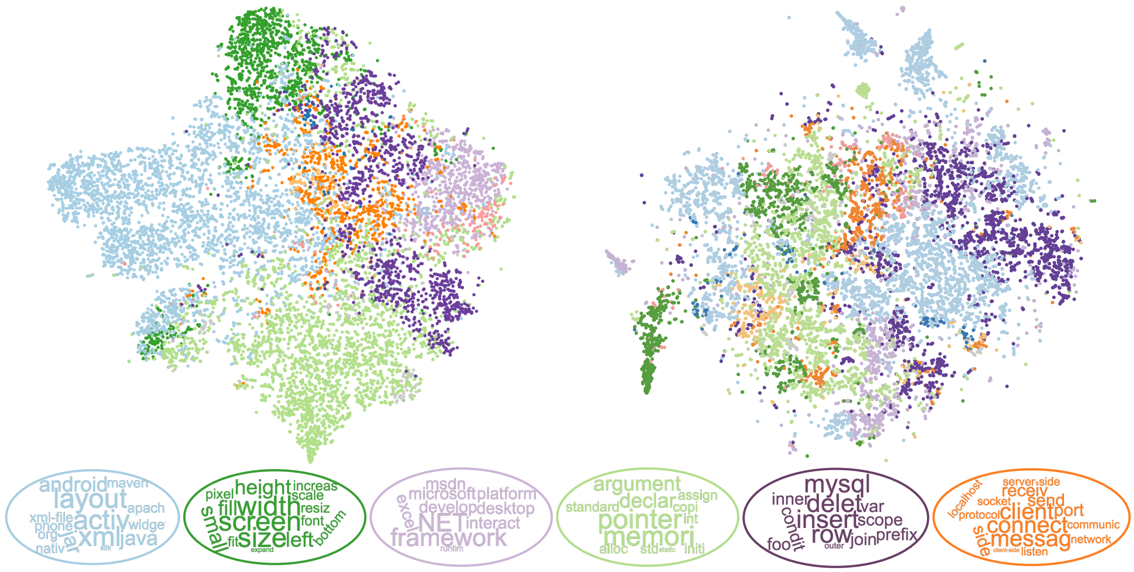

Figure 6.2: Visualization of user embedding with JNET (left) and TADW (right) and learnt topics in 2-D space of StackOverflow.

topics are most representative and meaningful at around 40 topics. Hence, we set topic number to 40 in the reported experiments. For the model parameters in JNET, we empirically set the precision for user embedding $\gamma = 0.1$, the precision for user-document topic proportion $\tau = 1$ and the variance for user affinity $\xi = 2$ and update other parameters in the M-step. The maximum number of iteration in our EM algorithm is set to 100. Both the source codes and data are available online [3].

6.3.2 The Learnt User Representation

We first study the quality of the learnt user representations from JNET. Users are mapped to 2-D space using the t-SNE package with the learnt user embeddings as input. For illustration purpose, we simply assign each user to the topic that he/she is closest to, i.e., $\text{argmax}(\Phi \cdot \boldsymbol{u}_i)$ and we mark users sharing the same interested topic with the same color. We also plot the most representative words of each topic learnt from JNET (i.e., $\text{argmax}\, p(w|\boldsymbol{\beta}_z)$), with the same color of the corresponding set of users. For comparison purpose, we also plot the learnt user representation from TADW. In order to assign the learnt user vectors from TADW to different topics, we adopted the same topic assignment method explained above by borrowing the learnt topic embeddings from JNET. And the visualizations of both methods with StakcOverflow dataset are shown in Figure 6.2.

[3] JNET. https://github.com/Linda-sunshine/JNET.

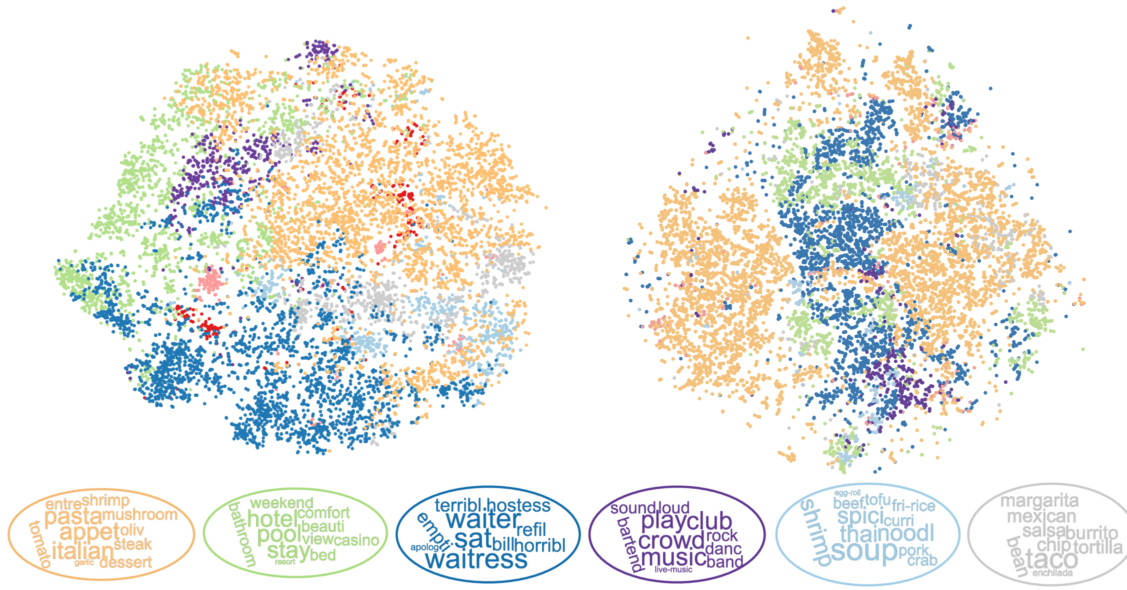

Figure 6.3: Visualization of user embedding with JNET (left) and TADW (right) and learnt topics in 2-D space of Yelp.

As we can find from the visualization of StackOverflow, users of similar interests are clearly clustered in the 2-D space, which indicates the descriptive power of our learnt user vectors. Meanwhile, representative topics are learnt as we can easily identify the theme of each word cloud, such as C++ (in light green circle), SQL (in dark purple circle) and java (in light blue circle) in the StackOverflow dataset. It is also interesting to find correlations among the users and topics by looking into their distances. The users in dark green are mainly interested in website development, thus are far away from the users who are interested in C++ (in light green). The users in orange care more about the network communication and they are overlapped with many other clusters of users focusing on SQL (in dark purple) and C++ (in light green) as network communication is an important component among different programming languages. However, the learnt user vectors from TADW are more scattered. For instance, users in light blue are splitted into four parts, indicating the relative low quality of the learnt user vectors.

We also provide the visualizations of Yelp dataset in Figure 6.3. Similar observations can also be found on Yelp dataset with JNET, for example, the users in light orange like Italian food and users in dark blue care about service. It is also interesting to find that users who like music and bar (in purple) are far away from the users who are fans of thai food (in light blue). With the learnt user vectors from TADW, users are scattered, such as users in light orange are clearly separated into two parts, which further verifies the decent quality of learnt user representations from JNET over TADW.

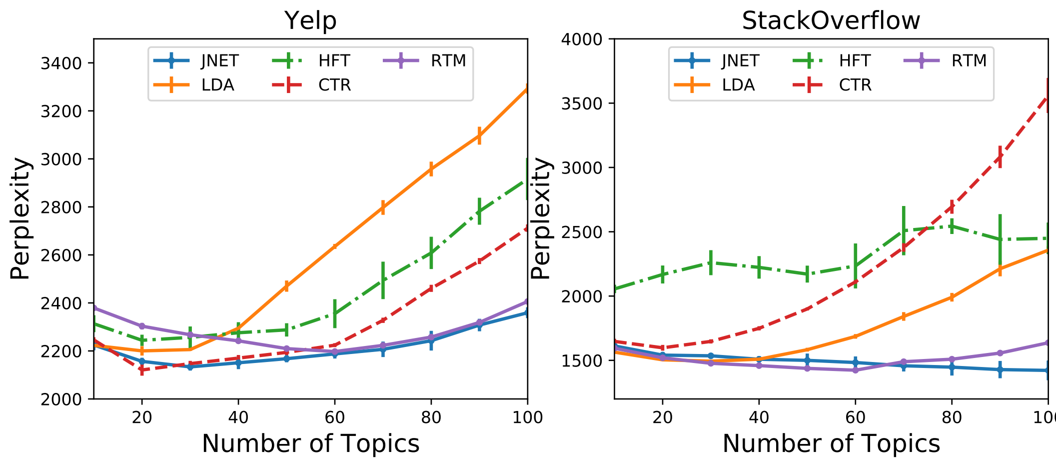

Figure 6.4: Perplexity comparison on Yelp and StackOverflow.

6.3.3 Document Modeling

In order to verify the predictive power of the proposed model, we first evaluated the generalization quality of JNET on the document modeling task. We compared all the topic model based solutions by their ***perplexity*** on a held-out test set. Formally, the perplexity for a set of held-out documents is calculated as follows [48]:

$$perplexity(D_{test}) = \exp\big(-\frac{\sum_{d\in D_{test}} \log p(\boldsymbol{w}_d)}{\sum_{d\in D_{test}} |d|}\big)$$

where $p(\boldsymbol{w}_d)$ is the likelihood of each held-out document given by a trained model. A lower perplexity indicates better generalization quality of a model.

Figure 6.4 reports the mean and variance of the perplexity for each model with 5-fold cross validation over different topic sizes. JNET achieved the best predictive power on the hold-out dataset, especially when an appropriate topic size is assigned. RTM achieved comparative performance as it utilizes the connectivity information among users, but it is limited by not being able to capture the variance within each user's different documents. The other baselines do not explicitly model network data, i.e., LDA, HFT and CTR, and therefore suffer in performance as they cannot take the advantage of network structure in modeling users.

A good joint modeling of network structure and text content should complement each other to facilitate a more effective user representation learning. Hence, we expect a good model to learn reasonable representations on users lacking text information, a.k.a., cold-start users, by utilizing network structure. We randomly selected 200 users and held out all their text content for testing.

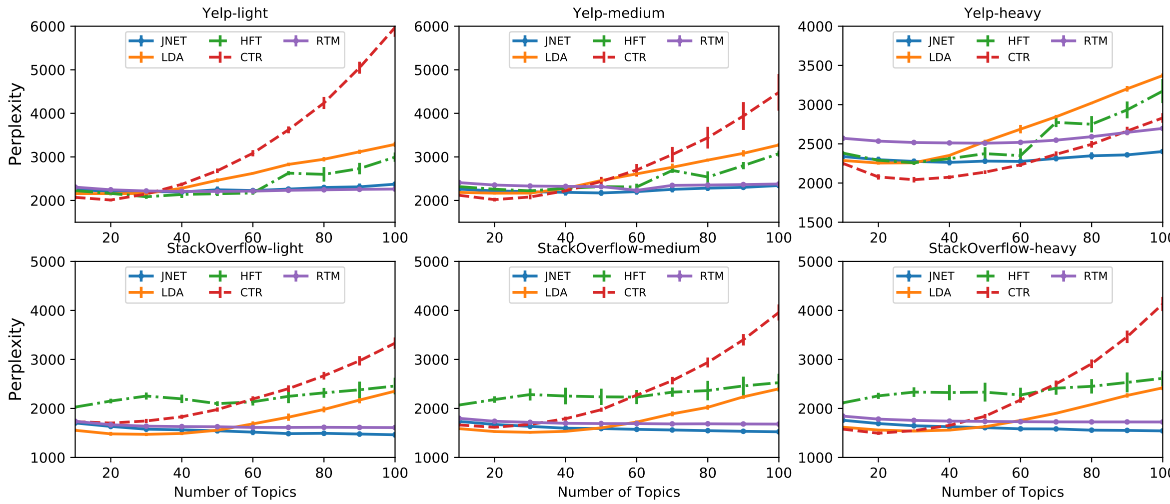

Figure 6.5: Comparison of perplexity in cold-start users on Yelp and StackOverflow.

Regarding to the number of social connections each testing user has in training data, we further consider three different sets of users, and name them as light, medium and heavy users, to give a finer analysis with respect to the degree of connectivity in cold-start setting. We end up with three sets of users and each set consists of 200 users. The threshold for selecting different sets of users is based on the statistics of each dataset. As a result, we selected 5 and 20 as the connectivity threshold for Yelp, 5 and 15 as the threshold for StackOverflow respectively. That is, in Yelp, light users have fewer than 5 friends, medium users have more than 5 friends while fewer than 20 friends and heavy users have more than 20 friends. We compared JNET against four baselines, i.e., LDA, HFT, RTM, CTR for evaluation purpose. We reported the perplexity on the held-out test documents regarding to the three sets of users, in Figure 6.5, respectively.

As we can observe in Figure 6.5, JNET performed consistently better on the testing documents for the three different sets of unseen users on Yelp dataset, which indicates the advantage of utilizing network information in addressing cold-start content prediction issue. The benefit of network is further verified across different sets of users as heavily connected users can achieve better performance improvement compared with text only user representation model, i.e., LDA. Similar conclusion is obtained for StackOverflow dataset, while we neglect it due to the space limit.

6.3.4 Link Prediction

The predictive power of JNET is not only reflected in unseen documents, but also in missing links. In the task of link prediction, the key component is to infer the similarity between users. We split

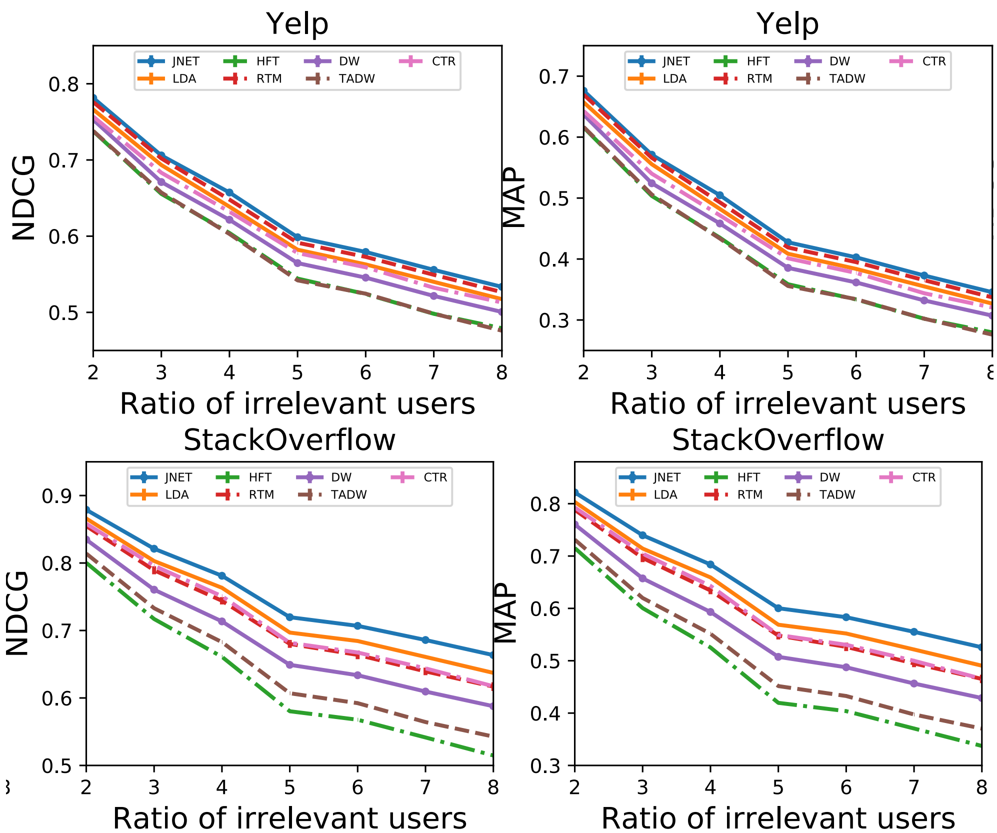

Figure 6.6: The performance comparison of link suggestion on Yelp and StackOverflow.

the observed social connections into 5 folds. Each time, we held out one fold of edges for testing and utilized the rest for model training, together with users' text content. In order to construct a valid set of ranking candidates for each testing user, we randomly injected irrelevant users (non-friends) for evaluation purpose. And the number of irrelevant users is proportional to the number of connections the testing user has, i.e., $t \times$ number of social connections. We rank users based on their cosine similarity to the target user. Normalized discounted cumulative gain (NDCG) and mean average precision (MAP) are used to measure the quality of ranking. We started with the ratio between irrelevant users and relevant users being $t = 2$ and increased the ratio to $t = 8$ to make the task more challenging to further verify the effectiveness of the learnt user representations.

To compare the prediction performance, we tested five baselines, i.e., LDA, HFT, RTM, DW and TADW. We reported the Normalized Discounted Cumulative Gain (NDCG) and Mean Average Precision (MAP) performance for the two datasets in Figure 6.6.

As we can find in the results, JNET achieved encouraging performance on both datasets, which indicates effective user representations are learnt to recover network structure. In Yelp dataset, network only solutions, i.e., DW, and text only solutions, i.e., LDA and HFT, cannot take the full

Table 6.1: The performance comparison of link prediction for light users of cold-start setting on StackOvewrflow.

Models	Light Ratio=2 NDCG/MAP	Light Ratio=4 NDCG/MAP	Light Ratio=6 NDCG/MAP
LDA	0.786/0.648	0.664/0.477	0.632/0.431
HFT	0.666/0.493	0.543/0.333	0.483/0.259
RTM	0.777/0.642	0.688/0.514	0.627/0.433
TADW	0.695/0.525	0.583/0.373	0.515/0.291
JNET	**0.794/0.664**	**0.697/0.534**	**0.643/0.453**

Table 6.2: The performance comparison of link prediction for medium users of cold-start setting on StackOvewrflow.

Models	Medium Ratio=2 NDCG/MAP	Medium Ratio=4 NDCG/MAP	Medium Ratio=6 NDCG/MAP
LDA	0.774/0.597	0.677/0.451	0.612/0.364
HFT	0.671/0.461	0.562/0.313	0.492/0.226
RTM	0.801/0.638	0.709/0.495	0.654/0.419
TADW	0.696/0.481	0.591/0.336	0.532/0.263
JNET	**0.812/0.649**	**0.724/0.511**	**0.663/0.425**

advantages of both modalities of user-generated data to capture user intents, while RTM achieved descent performance due to the integration of content and network modeling. Since the way of constructing network in StackOverflow is more content oriented, the performance of link suggestion on StackOverflow would prefer the text based solutions, which explains the comparable performance of LDA. Though TADW utilizes both modalities for user modeling, it fails to capture the dependency between them, leading to the poor performance on this task.

In practice, link prediction for unseen users is especially useful. For example, friend recommendation for new users in a system: they have very few or no friends, while they may associate with rich text content. This is also known as “cold-start” link prediction. Network only solutions will suffer from the lack of knowledge in such users. However, our proposed model can overcome this limitation by utilizing user-generated text content to learn representative user vectors, thus to provide helpful link prediction results.

In order to study the models’ predictive power in the cold-start setting, we randomly sampled three sets of users, regarding to the number of documents each user has, and name them as light, medium and heavy users. Each set of users consists of 200 users, and we selected 10 and 50 as the threshold for Yelp, 15 and 50 as the threshold for StackOverflow respectively. For example, in StackOverflow, light users have fewer than 15 posts, medium users have more than 15 but fewer than 50 posts, and heavy users have more than 50 posts. We compared the proposed model against four baselines, i.e., LDA, HFT, RTM and TADW for evaluation purpose. DW cannot learn representations for users

Table 6.3: The performance comparison of link prediction for heavy users of cold-start setting on StackOvewrflow.

Models	Heavy		
	Ratio=2	Ratio=4	Ratio=6
	NDCG/MAP	NDCG/MAP	NDCG/MAP
LDA	0.818/0.581	0.745/0.443	0.697/0.366
HFT	0.682/0.389	0.591/0.250	0.532/0.179
RTM	0.837/0.624	0.760/0.481	0.711/0.399
TADW	0.739/0.448	0.639/0.298	0.587/0.229
JNET	**0.842/0.626**	**0.763/0.483**	**0.713/0.399**

without any network information, thus is excluded in this experiment. We also randomly injected irrelevant users as introduced before for evaluation and we varied the ratio between irrelevant users and relevant users to change the difficulty of the task. We reported the NDCG and MAP performance on these three sets of users in Stackoverflow dataset with three differnet ratios, i.e., 2, 4 and 6, in Table 6.3, respectively.

JNET achieved consistently favorable performance in link prediction in users without any social connections, as accurate proximity between users are properly identified with its user representations learnt from text data. Comparing across groups, better performance is achieved for users with more text documents. Similar results were obtained on Yelp dataset as well, but omitted due to space limit.

6.3.5 Expert Recommendation

In the sampled StackOverflow dataset, the average number of answers for questions is as low as 1.14, which indicates the difficulty for getting an expert to answer the question. If the system can suggest the right user to answer the posted questions, e.g., push the question to the selected user, more questions would be answered more quickly and accurately. We conjecture the learnt topic distribution of each question in StackOverflow, together with the identified user representation, facilitate the task of expert recommendation for question answering. The task can be further decomposed into two components: whether the question falls in a user's skill set; and whether the user who asked question shares similar interests with the potential candidate expert. With the learnt topic embeddings Φ and each user's embedding u_i, each user's interest over the topics can be characterized as a mapping from the topic embeddings to the user's embedding, i.e., $\Phi \cdot u_i$. Together with the learnt topic distribution of each question, we can estimate the proximity between the question and the user's expertise to score the alignment between the question and the user. In the meanwhile, the closeness between

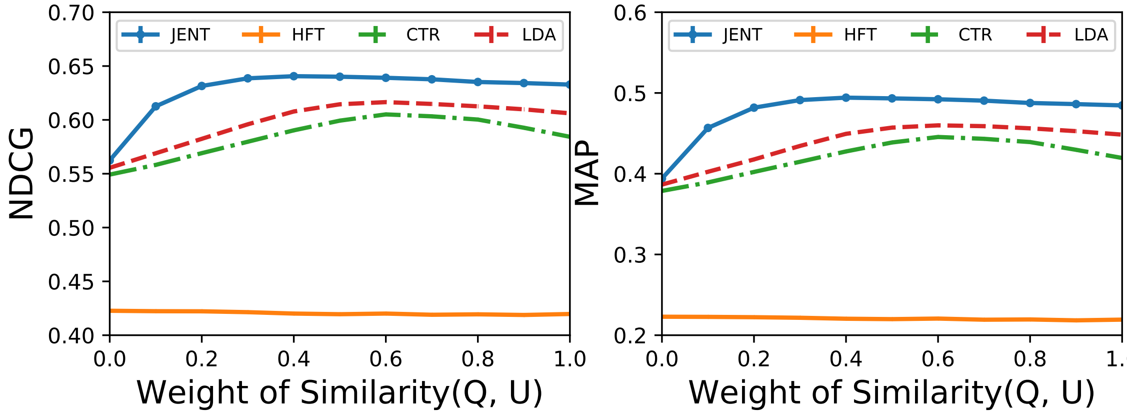

Figure 6.7: Expert recommendation on StackOverflow.

users can be simply measured by the distance of their corresponding embedding vectors. As a result, the task can be formalized as finding the user that achieves the highest relatedness with the given question, where we define the relatedness as follows:

$$\text{score} = \alpha \cdot \text{cosine}(\boldsymbol{u}_i \cdot \Phi, \boldsymbol{\theta}_{id}) + (1 - \alpha) \cdot \text{cosine}(\boldsymbol{u}_i, \boldsymbol{u}_j) \tag{6.3.1}$$

Due to the limited number of answers for each question in our dataset, we selected 1,816 questions with more than 2 answers for the experiment. Besides the users that answered the given question, we also incorporated irrelevant users for each question for evaluation purpose. And the number of irrelevant users is 10 times of the number of answers. We compared against the learnt topic distributions of questions and user representations from LDA, HFT and CTR. As we tune the weight between the two components in Eq (6.3.1), we plot the corresponding NDCG and MAP in Figure 6.7.

The proposed model achieved very promising performance in the recommendation task, as it explicitly models user's expertise and the given question in the topic space. The estimated similarities between user-user and user-content accurately align the question to the right user. However, baseline models can only capture the similarity between questions and users based on their topical similarity, which is insufficient in this task. Interestingly, as we gradually increased the weight of question-content similarity from 0, JNET's performance peaked, which indicates the relative importance between user-user and user-content similarities for this specific problem.

6.4 Conclusion

In the work, we captured user intents via user representation learning by explicitly modeling the structural dependency among different modalities of user-generated data. We proposed a complete generative model to integrate user representation learning with content modeling and social network modeling. By constructing a shared latent space to embed both users and topics, the relationship between different modalities can be clearly depicted, together with users' preferences towards different objects. The learnt user representations are interpretable and predictive, indicated by the performance improvement in many important tasks such as link suggestions and expert finding.

Chapter 7

Conclusions and Future Works

In the era of Internet, people interact with diverse information service systems extensively everyday, such as search engines and social media website, to satisfy their needs and desires. In order to provide the service satisfying user needs and preferences quickly and efficiently, computational user modeling becomes an essential component which enables an in-depth understanding of user intents. The design of these computational user models introduced in the dissertation gets inspired from ***social psychology principles***, such as imposing certain assumptions and quantifying concepts, which in turn provides effective computational techniques for research in social psychology. We try to ***bridge the gap between social psychology and computational user behavior modeling*** systematically, which make contributions to both communities.

In this dissertation, we focus on exploring multiple modalities of user-generated data to capture diverse user intents via computational user modeling. We start from exploring user-generated text content to examine their distinct ways of expressing attitudes. Multi-task learning is utilized to encode the task relatedness, to build the connectivity among users. In addition to the implicit connectivity, the availability of network structure provides the opportunity to encode task relatedness. Thus, it is further incorporated to achieve a comprehensive understanding of user intents through the learning of user representations.

7.1 Conclusions

In the dissertation, we proposed a multi-modal user intent learning framework via performing computational user modeling. In particular, we solved the user intent learning problem raised in the introduction from two different perspectives.

• **Modeling Opinionated Text for Personalized Sentiment Analysis.** In the first part of the dissertation, we mainly analyze users' opinionated text to understand their different ways of expressing opinions. The proposed MTLinAdapt and cLinAdapt models achieve personalized sentiment analysis effectively, especially for the users with limited amount of textual information. Within the specific task of opinion mining, the two approaches learn effective user profiles with respect to their preferences of opinionated words selected for expressing attitudes, i.e., the set of weights for the corresponding opinion words. More specifically, MTLinAdapt performs the user profile learning from each individual's opinionated reviews to capture the nuance in expressing attitudes. By realizing the clustering property among users, cLinAdapt further imposes a non-parametric Dirichlet Process prior over users' personalized models to learn users' diverse opinions in a group manner, to better alleviate the data sparsity issue and enable the implicit connectivity among users. With the limit of exploring text content only, the user profiles mainly characterize users' sentiment and may not generalize well to other tasks.

• **Incorporating Network for Holisitic User Behavior Modeling.** In the second part of the dissertation, we further incorporate the available network structure to achieve a more comprehensive understanding of user intents. The proposed HUB model performs holistic user behavior modeling by integrating companion tasks of content modeling and network structure modeling. The learned user representation is a more general depiction of user preferences, which can facilitate many other applications such as friend recommendation. Limited by the implicit modeling of dependency between different modalities of user-generated data, HUB cannot provide a clear picture explaining the correlation between different modalities of user-generated data, i.e., the correlation between text content and network structure. Thus, JNET model resolves the concern by learning a shared low-dimensional space to embed different modalities of user-generated data so as to encode the relatedness and dependency among them. With the learnt user embeddings and topic embeddings available in the shared latent space, the closeness between textual content and network can be easily measured, which provides a more clear and precise user understanding via such decomposition so as to further facilitate the content recommendation.

7.2 Future Work

• **Incorporating Heterogeneous Data for User Representation Learning.** With more heterogeneous user-generated data available today, such as images posted in image-sharing platforms or short videos posted on video-sharing platforms, we have new resources to further explore user desires or needs due to the unique properties of these data, which can help achieve user understanding in a complementary way. For instance, personal images are growing explosively with the popularity of social media, which largely exhibit users' opinions, interests and emotions. Mining user intents from personal images itself can facilitate a number of applications, such as interest based community detection, image recommendation and advertising. Moreover, images can complement the other modalities to gain comprehensive user intent understanding. Compared with texts, images are more natural to express users interests and emotion as they usually conveys topics and themes. By properly analyzing posted images, they can either confirm the intent discovered by text or complement text to further exploit user intent. Besides, each individual user may have very limited amount of connections online while images help reveal the implicit connections among users. Indeed, images can help build interest graph among users as users who like, share or forward the same image are very likely to share the same interest, thus serving as great resources to help establish connections, or even construct the communities. In the framework of user intent learning, the utilization of heterogeneous data will definitely help overcome the data sparsity issue, thus to achieve more effective and accurate user representation learning, covering more aspects of each individual user.

• **Studying the Dimension of Time for Capturing Dynamics.** The aforementioned works introduced in the dissertation assume the models are static in analyzing user behaviors, thus are unable to capture the dynamic changes either individually or globally. The dimension of time provides us a different angle to interpret user behaviors, enabling numerous time-sensitive applications. A straightforward example would be users' social intent may evolve obviously over time. In their 20s to 30s, they care more about like-minded individuals or romantic partners; Later on, they may spend more time looking for cooperators and collaborators for successful career development, which cannot be captured by static user model clearly. More interestingly, the dimension of time enables us to identify users' long-term interests and short-term interest efficiently.

Beside the individual-level changes, the whole society may evolve over time. With the quick development of Internet, tremendous new information is emerging everyday, leading to a large amount of new concepts, either replacing the old ones or creating new concepts. The dynamic view can help exploit the evolution of culture, technology, economy and many other fields, resulting in a lot interesting research problems. The evolution of network helps understand the establishment of user connections and the formalization of communities, enabling vast amounts of applications such as friend recommendation, community suggestion and so on. At the same time, it also serves as great reference for social psychologists to study human behaviors in real world.

• **Multi-role User Representation Learning.** In practice, users usually participate in different communities and behave in certain ways accordingly, such as colleagues, family members, friends. This kind of community is not limited to explicit social circle, which can also be an implicit context. That is, users are usually associated with multiple roles under different contexts, leading one single user embedding insufficient to represent and calling for the learning of multi-role user representation.

The identification of such role information can provide a clear picture of the decomposition of all the interactions, which helps gain user understanding from a high-resolution. With such role identification available, a lot of applications can be enabled. For example, the specific role information can help find the candidate user or item for recommendation, and it also serve as the explanation to make the recommendation more concrete and convincing at the same time.

• **Scalability.** The large amounts of user-generated data brings in opportunity for achieving user understanding, while it also imposes challenges in the computational perspective. Especially, the challenges lie in two folds: the first is caused by the complex interactions among different modalities of user-generated data; and the second is raised by the network structure due to the pairwise property. Therefore, improving the scalability of the learning models can speed up the training process, imposing tremendous practical values directly. Especially, if we incorporate more modalities of user-generated data, learning compact user representations quickly and accurately would be challenging and difficult. We may comprise between speed and accuracy, to achieve a perfect balance for practical applications.

7.3 Broader Impacts

This research is an amalgamation of important aspects of both social psychology and computational modeling. The main aim of this work is two-fold, leverage social psychology principles to better design user behavior models as psychical world provides good references for virtual world; the computational user behavior modeling can in turn benefit the community of social psychology as it provides alternatives for researchers in social psychology to perform experiments instead of traditional surveys and polls. Though users' online behaviors might differ from their behaviors in psychical world, still, the principles in social psychology provide insights in designing computational models, such as motivating us to impose certain assumptions in performing user modeling. Since the data collection is an important component in the study of social psychology, the massive user-generated data naturally serves as great resources for this goal. In order to utilize the data, it is necessary to leverage computational techniques to mine the data, so as to understand user behaviors efficiently.

Currently, the Internet has infiltrated every aspect of our lives, producing large amounts of data everyday to facilitate the understanding of user behaviors online. Consequently, it will bring in billions of dollars of value to the economy and a huge amount of social capital to society. Though the systems are growing rapidly, they are still in their infancy, lacking rigorous principles and governance. Different from the real world, it is more difficult and challenging to monitor and govern online users. In the real world, interactions among individuals usually occur in a more direct way, i.e., face-to-face communication or interaction, which makes the intent understanding relatively easy, thus to perform regulation and governance effectively. However, online users usually behave anonymously, such as expressing opinions or making connections, which makes it hard to infer their intents, thus to regulate user behaviors timely and accurately. Due to the unique properties of online environment, a lot of problems emerge gradually. For instance, rumors spread on the Internet quickly and extensively, rile up users' emotions and moods, cause anxiety and bring in negative effectiveness. Terrorists and racists post evil comments online, which makes people uncomfortable and jittery. Also, cyber-bullying happens frequently in various forms, and harms both children and adults. Different from schoolyard bullying, teachers can't intervene on the Internet to protect children, which makes it a more difficult issue.

With these concerns, it is necessary to call for a healthy virtual social systems via establishing rigorous principles or performing certain regulations. And the computational user modeling has great impacts on building such healthy online social systems. By collecting the user-specific data, we can get to know user intents via proper computational modeling. Once the harmful intent is detected, corresponding actions should be performed to regulate so as to protect the online environment. For instance, when rumors are detected by the system, the information source should be cut off immediately to prevent diffusion and the person who spread the rumors should be punished and reeducated to avoid recommit. When racists are found online, the relevant comments and posts should be deleted and the person should be punished and reeducated. Not only the individual-level intent understanding can help build online social systems, global-level understanding also contributes to the establishment of online social systems. For instance, analyzing users' text content helps capture new Internet culture, thus to form the standard of Internet language.

The process acts as a social eco-system, we will ***1) gain knowledge of online users gradually via the exploration of massive user-generated data***, from which ***2) smart systems can be designed to detect abnormal or harmful behaviors such as cyber-bulling, so as to build healthy online social systems***. The sound social system will grow steadily and quickly, creating more powerful clues for inferring user understanding on both individual level and global level. Though we can only start with shallow understanding of user behaviors, it can still help detect abnormal user behaviors to regulate the online social system. With more sound systems established, users will participate in diverse activities with more interactions, generating more data for examination. The aforementioned two steps will repeat to establish mature and healthy online social system mutually. With perfect online social system, we can further integrate the online social system and physical social system to make a unified social system.

Bibliography

[1] 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.

[2] Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD*, pages 783–792. ACM, 2010.

[3] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th ACM SIGIR*, pages 83–92. ACM, 2014.

[4] Y-C Zhang, Matus Medo, Jie Ren, Tao Zhou, Tao Li, and Fan Yang. Recommendation model based on opinion diffusion. *EPL (Europhysics Letters)*, 80(6):68003, 2007.

[5] Mohammad-Ali Abbasi, Sun-Ki Chai, Huan Liu, and Kiran Sagoo. Real-world behavior analysis through a social media lens. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 18–26. Springer, 2012.

[6] 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.

[7] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM CIKM*, pages 824–831. ACM, 2005.

[8] Fernando Rivera-Illingworth, Victor Callaghan, and Hani Hagras. A connectionist embedded agent approach for abnormal behaviour detection in intelligent health care environments. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 4, pages 3565–3570. IEEE, 2004.

[9] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

[10] Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. Risk prediction on electronic health records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1910–1919. ACM, 2018.

[11] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 31–40. ACM, 2010.

[12] Qingyao Ai, Yongfeng Zhang, Keping Bi, Xu Chen, and W Bruce Croft. Learning a hierarchical embedding model for personalized product search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 645–654. ACM, 2017.

[13] Hongning Wang, ChengXiang Zhai, Feng Liang, Anlei Dong, and Yi Chang. User modeling in search logs via a nonparametric bayesian approach. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 203–212. ACM, 2014.

[14] Eren Manavoglu, Dmitry Pavlov, and C Lee Giles. Probabilistic user behavior models. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 203–210. IEEE, 2003.

[15] Chunfeng Yang, Huan Yan, Donghan Yu, Yong Li, and Dah Ming Chiu. Multi-site user behavior modeling and its application in video recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 175–184. ACM, 2017.

[16] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD*, pages 701–710. ACM, 2014.

[17] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th WWW*, pages 1067–1077, 2015.

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

[19] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010.

[20] 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.

[21] Jiliang Tang, Yi Chang, and Huan Liu. Mining social media with social theories: A survey. *ACM SIGKDD Explorations Newsletter*, 15(2):20–29, 2014.

[22] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, 58(7):1019–1031, 2007.

[23] 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.

[24] Lin Gong, Mohammad Al Boni, and Hongning Wang. Modeling social norms evolution for personalized sentiment classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 855–865, 2016.

[25] Bertram F Malle and Joshua Knobe. The folk concept of intentionality. *Journal of experimental social psychology*, 33(2):101–121, 1997.

[26] 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.

[27] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[28] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[29] Wenliang Gao, Nobuhiro Kaji, Naoki Yoshinaga, and Masaru Kitsuregawa. Collective sentiment classification based on user leniency and product popularity. , 21(3):541–561, 2014.

[30] Ivan Titov and Ryan T McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, volume 8, pages 308–316. Citeseer, 2008.

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

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

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

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

[35] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

[36] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[37] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI*, pages 1361–1370. ACM, 2010.

[38] Prescott Lecky. Self-consistency; a theory of personality. 1945.

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

[40] 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.

[41] Lin Gong, Benjamin Haines, and Hongning Wang. Clustered model adaption for personalized sentiment analysis. In *Proceedings of the 26th WWW*, pages 937–946, 2017.

[42] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[43] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD*, pages 168–177. ACM, 2004.

[44] Sanjiv Das and Mike Chen. Yahoo! for amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific finance association annual conference (APFA)*, volume 35, page 43. Bangkok, Thailand, 2001.

[45] Alison Huettner and Pero Subasic. Fuzzy typing for document management. *ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes*, pages 26–27, 2000.

[46] Benjamin Snyder and Regina Barzilay. Multiple aspect ranking using the good grief algorithm. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 300–307, 2007.

[47] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[48] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[49] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on UAI*, pages 487–494. AUAI Press, 2004.

[50] Chong Wang and David M Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD*, pages 448–456. ACM, 2011.

[51] Max Woolf. A statistical analysis of 1.2 million amazon reviews. `http://minimaxir.com/2014/06/reviewing-reviews/`, 2014.

[52] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pages 607–618. ACM, 2013.

[53] Anindya Ghose, Panagiotis G Ipeirotis, and Beibei Li. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science*, 31(3):493–520, 2012.

[54] Evelien Otte and Ronald Rousseau. Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, 28(6):441–453, 2002.

[55] Jennifer Golbeck. *Analyzing the social web.* Newnes, 2013.

[56] Emily M Jin, Michelle Girvan, and Mark EJ Newman. Structure of growing social networks. *Physical review E*, 64(4):046132, 2001.

[57] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P Gummadi. Measuring user influence in twitter: The million follower fallacy. In *fourth international AAAI conference on weblogs and social media*, 2010.

[58] Zepeng Huo, Xiao Huang, and Xia Hu. Link prediction with personalized social influence. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[59] Hanghang Tong, Christos Faloutsos, Christos Faloutsos, and Yehuda Koren. Fast direction-aware proximity for graph mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 747–756. ACM, 2007.

[60] Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 635–644. ACM, 2011.

[61] Yuchung J Wang and George Y Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

[62] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[63] Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):43, 2013.

[64] Jaewon Yang, Julian McAuley, and Jure Leskovec. Community detection in networks with node attributes. In *Data Mining (ICDM), 2013 IEEE 13th international conference on*, pages 1151–1156. IEEE, 2013.

[65] Simon Bourigault, Cedric Lagnier, Sylvain Lamprier, Ludovic Denoyer, and Patrick Gallinari. Learning social network embeddings for predicting information diffusion. In *Proceedings of the 7th ACM WSDM*, pages 393–402. ACM, 2014.

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

[67] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

[68] Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd ACM SIGKDD*, pages 1105–1114. ACM, 2016.

[69] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *AAAI*, pages 203–209, 2017.

[70] Lei Tang and Huan Liu. Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD*, pages 817–826. ACM, 2009.

[71] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234. ACM, 2016.

[72] Quanjun Chen, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki. Learning deep representation from big and heterogeneous data for traffic accident inference. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[73] Dingyuan Zhu, Peng Cui, Daixin Wang, and Wenwu Zhu. Deep variational network embedding in wasserstein space. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2827–2836. ACM, 2018.

[74] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63. Association for Computational Linguistics, 2011.

[75] 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.

[76] Kewei Cheng, Jundong Li, Jiliang Tang, and Huan Liu. Unsupervised sentiment analysis with signed social networks. In *AAAI*, pages 3429–3435, 2017.

[77] Jiliang Tang, Chikashi Nobata, Anlei Dong, Yi Chang, and Huan Liu. Propagation-based sentiment analysis for microblogging data. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 577–585. SIAM, 2015.

[78] Federico Alberto Pozzi, Daniele Maccagnola, Elisabetta Fersini, and Enza Messina. Enhance user-level sentiment analysis on microblogs with approval relations. In *Congress of the Italian Association for Artificial Intelligence*, pages 133–144. Springer, 2013.

[79] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. Network representation learning with rich text information. In *IJCAI*, pages 2111–2117, 2015.

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

[81] Theodoros Evgeniou, Charles A Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. In *Journal of Machine Learning Research*, pages 615–637, 2005.

[82] 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.

[83] Tony Jebara. Multi-task feature and kernel selection for svms. In *Proceedings of the twenty-first international conference on Machine learning*, page 55. ACM, 2004.

[84] Antonio Torralba, Kevin P Murphy, William T Freeman, et al. Sharing features: efficient boosting procedures for multiclass object detection. *CVPR (2)*, 3, 2004.

[85] Kai Yu, Volker Tresp, and Anton Schwaighofer. Learning gaussian processes from multiple tasks. In *Proceedings of the 22nd International Conference on Machine Learning (ICML-05)*, pages 1012–1019, 2005.

[86] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

[87] Jonathan Baxter. A model of inductive bias learning. *J. Artif. Intell. Res.(JAIR)*, 12:149–198, 2000.

[88] Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003.

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

[90] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[91] A Evgeniou and Massimiliano Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19:41, 2007.

[92] Hongliang Fei, Ruoyi Jiang, Yuhao Yang, Bo Luo, and Jun Huan. Content based social behavior prediction: a multi-task learning approach. In *Proceedings of the 20th ACM CIKM*, pages 995–1000. ACM, 2011.

[93] Shelley E. Taylor, Letitia Anne. Peplau, and David O. Sears. *Social psychology*. Pearson/Prentice Hall, 2006.

[94] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[95] Donnel A Briley, Michael W Morris, and Itamar Simonson. Reasons as carriers of culture: Dynamic versus dispositional models of cultural influence on decision making. *Journal of consumer research*, 27(2):157–178, 2000.

[96] 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.

[97] Paul R Ehrlich and Simon A Levin. The evolution of norms. *PLoS Biol*, 3(6):e194, 2005.

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

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

[100] 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.

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

[102] 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.

[103] Susan Shott. Emotion and social life: A symbolic interactionist analysis. *American journal of Sociology*, pages 1317–1334, 1979.

[104] Arlie Russell Hochschild. The sociology of feeling and emotion: Selected possibilities. *Sociological Inquiry*, 45(2-3):280–307, 1975.

[105] Elinor Ostrom. Collective action and the evolution of social norms. *Journal of Natural Resources Policy Research*, 6(4):235–252, 2014.

[106] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

[107] 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.

[108] Yelp. Yelp dataset challenge. `https://www.yelp.com/dataset_challenge`, 2016.

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

[110] Henry Brighton and Chris Mellish. Advances in instance selection for instance-based learning algorithms. *Data mining and knowledge discovery*, 6(2):153–172, 2002.

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

[112] Krzysztof C Kiwiel. Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical programming*, 90(1):1–25, 2001.

[113] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260. ACM, 2002.

[114] 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.

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

[116] 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.

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

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

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

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

[121] 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.

[122] 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.

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

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

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

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

[127] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th international conference on World wide web*, pages 261–270. ACM, 2009.

[128] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[129] John C Turner. Social categorization and the self-concept: A social cognitive theory of group behavior. *Advances in group processes*, 2:77–122, 1985.

[130] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392, 2005.

[131] Rina S Onorato and John C Turner. Fluidity in the self-concept: the shift from personal to social identity. *European Journal of Social Psychology*, 34(3):257–278, 2004.

[132] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

[133] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *IEEE ICDM*, pages 263–272. Ieee, 2008.

[134] Steffen Rendle. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE, 2010.

[135] Gerhard Fischer. User modeling in human–computer interaction. *User modeling and user-adapted interaction*, 11(1-2):65–86, 2001.

[136] Alfred Kobsa. Generic user modeling systems. *User modeling and user-adapted interaction*, 11(1-2):49–63, 2001.

[137] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD*, pages 137–146. ACM, 2003.

[138] Lin Gong and Hongning Wang. When sentiment analysis meets social network: A holistic user behavior modeling in opinionated data. In *Proceedings of the 24th ACM SIGKDD*, pages 1455–1464. ACM, 2018.

[139] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[140] Hongbo Deng, Jiawei Han, Hao Li, Heng Ji, Hongning Wang, and Yue Lu. Exploring and inferring user–user pseudo-friendship for sentiment analysis with heterogeneous networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7(4):308–321, 2014.

[141] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[142] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.

[143] Michelle Rae Tuckey and Neil Brewer. The influence of schemas, stimulus ambiguity, and interview schedule on eyewitness memory over time. *Journal of Experimental Psychology: Applied*, 9(2):101, 2003.

[144] Jonathan Chang and David Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, pages 81–88, 2009.

[145] David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[146] Robert B Cialdini and Melanie R Trost. Social influence: Social norms, conformity and compliance. 1998.

[147] Olivier Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1097–1105, New York, NY, USA, 2014. ACM.

[148] John S Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th UAI*, pages 43–52, 1998.

[149] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[150] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[151] Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino, and Belle L Tseng. On evolutionary spectral clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(4):17, 2009.

[152] Peter Willett. The porter stemming algorithm: then and now. *Program*, 40(3):219–223, 2006.

[153] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Smart stopword list, 2004.

[154] Arvind Agarwal and Saurabh Kataria. Multitask learning for sequence labeling tasks. *arXiv preprint arXiv:1404.6580*, 2014.

[155] Daniel Beck, Trevor Cohn, and Lucia Specia. Joint emotion analysis via multi-task gaussian processes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1798–1803. ACL, 2014.

[156] Trevor Cohn and Lucia Specia. Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In *ACL (1)*, pages 32–42. Citeseer, 2013.

[157] Pedro Henrique Calais Guerra, Adriano Veloso, Wagner Meira Jr, and Virgílio Almeida. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM, 2011.

[158] Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 210–219. ACM, 2007.

[159] Rajat Raina, Andrew Y Ng, and Daphne Koller. Constructing informative priors using transfer learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 713–720. ACM, 2006.

[160] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics, 2011.

[161] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics, 2010.

[162] Serge Moscovici, Carol Sherrard, and Greta Heinz. *Social influence and social change*, volume 10. Academic Press London, 1976.

[163] Letitia Anne Peplau Taylor, Shelley E. and David O. Sears. *Social Psychology.* Pearson Education New Jersey, 2006.

[164] Mark Dredze and Koby Crammer. Online methods for multi-domain learning and adaptation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 689–697. Association for Computational Linguistics, 2008.

[165] Shu Huang, Wei Peng, Jingxuan Li, and Dongwon Lee. Sentiment and topic analysis on social media: a multi-task multi-label classification approach. In *Proceedings of the 5th annual ACM web science conference*, pages 172–181. ACM, 2013.

[166] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[167] W Glynn Mangold and David J Faulds. Social media: The new hybrid element of the promotion mix. *Business horizons*, 52(4):357–365, 2009.

[168] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[169] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM, 2005.

[170] Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

[171] Johan Bollen, Alberto Pepe, and Huina Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *arXiv preprint arXiv:0911.1583*, 2009.

[172] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *ICWSM*, 2011.

[173] Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th WWW*, pages 519–528. ACM, 2003.

[174] Niklas Jakob, Stefan Hagen Weber, Mark Christoph Müller, and Iryna Gurevych. Beyond the stars: exploiting free-text user reviews to improve the accuracy of movie recommendations. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 57–64. ACM, 2009.

[175] Yue Lu and Chengxiang Zhai. Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th international conference on World Wide Web*, pages 121–130. ACM, 2008.

[176] Siamak Faridani. Using canonical correlation analysis for generalized sentiment analysis, product recommendation and search. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 355–358. ACM, 2011.

[177] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th COLING*, pages 1367–1373. ACL, 2004.

[178] Michael J Pazzani. A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5-6):393–408, 1999.

[179] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407, 1990.

[180] John D Mayer, Richard D Roberts, and Sigal G Barsade. Human abilities: Emotional intelligence. *Annu. Rev. Psychol.*, 59:507–536, 2008.

[181] Timothy La Fond and Jennifer Neville. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th WWW*, pages 601–610. ACM, 2010.

[182] James S Coleman and James Samuel Coleman. *Foundations of social theory*. Harvard university press, 1994.

[183] Albert Bandura. Social foundations of thought and action: A social cognitive perspective. *Englewood Cliffs, NJ: Princeton-Hall*, 1986.

[184] Cheryl L Perry, Tom Baranowski, and Guy S Parcel. How individuals, environments, and health behavior interact: social learning theory. 1990.

[185] Raya Fidel and Michael Crandall. Users' perception of the performance of a filtering system. In *ACM SIGIR Forum*, volume 31, pages 198–205. ACM, 1997.

[186] Bandura Albert. Social foundations of thought and action: A social cognitive theory. *NY.: Prentice-Hall*, 1986.

[187] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *NIPS*, pages 1385–1392, 2004.

[188] 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.

[189] Laura Frances Bright. *Consumer control and customization in online environments: An investigation into the psychology of consumer choice and its impact on media enjoyment, attitude and behavioral intention*. The University of Texas at Austin, 2008.

[190] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[191] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[192] Michael A Hogg and Kipling D Williams. From i to we: Social identity and the collective self. *Group dynamics: Theory, research, and practice*, 4(1):81, 2000.

[193] Marilynn B Brewer. The social self: On being the same and different at the same time. *Personality and social psychology bulletin*, 17(5):475–482, 1991.

[194] Andrew B Goldberg and Xiaojin Zhu. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52. Association for Computational Linguistics, 2006.

[195] Michael A Hogg and Scott Tindale. *Blackwell handbook of social psychology: Group processes*. John Wiley & Sons, 2008.

[196] Marilynn B Brewer. Optimal distinctiveness, social identity, and the self. 2003.

[197] Jörn Davidsen, Holger Ebel, and Stefan Bornholdt. Emergence of a small world from local interactions: Modeling acquaintance networks. *Physical Review Letters*, 88(12):128701, 2002.

[198] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[199] Michael Wooldridge and Nicholas R Jennings. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995.

[200] Shlomo Berkovsky, Tsvi Kuflik, and Francesco Ricci. Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, 18(3):245–286, 2008.

[201] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(2009):12, 2009.

[202] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM, 2010.

[203] Antonio Reyes, Paolo Rosso, and Tony Veale. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268, 2013.

[204] David Burth Kurka, Alan Godoy, and Fernando J Von Zuben. Online social network analysis: A survey of research applications in computer science. *arXiv preprint arXiv:1504.05655*, 2015.

[205] Michael A Hogg. Social categorization, depersonalization, and group behavior. *Blackwell handbook of social psychology: Group processes*, 4:56–85, 2001.

[206] Tom AB Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100, 1997.

[207] Charles Kemp, Thomas L Griffiths, and Joshua B Tenenbaum. Discovering latent classes in relational data. 2004.

[208] Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5, 2006.

[209] Amr Ahmed, Yucheng Low, Mohamed Aly, Vanja Josifovski, and Alexander J Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 114–122. ACM, 2011.

[210] Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26. ACM, 2006.

[211] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

[212] Fangzhao Wu, Yongfeng Huang, and Jun Yan. Active sentiment domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1701–1711, 2017.

[213] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[214] Yuchen Zhang, Weizhu Chen, Dong Wang, and Qiang Yang. User-click modeling for understanding and predicting search-behavior. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1388–1396. ACM, 2011.

[215] Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *science*, 311(5757):88–90, 2006.

[216] Robert B Allen. User models: theory, method, and practice. *International Journal of man-machine Studies*, 32(5):511–543, 1990.

[217] Kurt Koffka. *Principles of Gestalt psychology.* Routledge, 2013.

[218] Paul DiMaggio. Culture and cognition. *Annual review of sociology*, 23(1):263–287, 1997.

[219] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *Proceedings of the 17th WWW*, pages 101–110. ACM, 2008.

[220] Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, and David M Blei. Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems*, pages 17–24, 2004.

[221] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[222] Ryen W White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22nd WWW*, pages 1411–1420. ACM, 2013.

[223] Duyu Tang, Bing Qin, and Ting Liu. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of the 53rd ACL*, volume 1, pages 1014–1023, 2015.

[224] Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 EMNLP*, pages 1650–1659, 2016.

[225] Alfred V. Aho and Jeffrey D. Ullman. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ, 1972.

[226] American Psychological Association. *Publications Manual.* American Psychological Association, Washington, DC, 1983.

[227] Association for Computing Machinery. In *Computing Reviews*, volume 24, pages 503–512. 1983.

[228] Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133, 1981.

[229] Dan Gusfield. *Algorithms on Strings, Trees and Sequences.* Cambridge University Press, Cambridge, UK, 1997.

[230] TM Heskes. Empirical bayes for learning to learn. 2000.

[231] Sebastian Thrun and Lorien Pratt. *Learning to learn.* Springer Science & Business Media, 2012.