WELLNESS PROFILING ON SOCIAL NETWORKS

MOHAMMAD AKBARI

(M.Sc.(Hons.), Amirkabir University of Technology)

(B.Eng.(Hons.), Kharazmi University)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR INTEGRATIVE SCIENCES AND ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2016

Declaration

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Mohammad Akbari

July 2016

Acknowledgments

This thesis would not have been possible without the support of many people. I would like to express my sincere gratitude to my adviser, Prof. Tat-Seng Chua, for the continuous support of my PhD study and research, for his patience, motivation, immense knowledge, and above all, his honest and serious behavior in conducting research. His guidance helped me throughout my research and writing of this thesis.

I would like to express my deepest appreciation to my Thesis Advisory Committee, Prof. Mohan Kankanhalli and Prof. Michael S. Brown, for their support and guidance during my PhD journey. I would also like to thank Prof. Xia Hu who guided me during my PhD study.

I thank my labmates in the Lab for Media Search (LMS) for their friendships, for the times we had together, for the weekly meetings we had to discuss our research problems, and for all the fun we had during my PhD study. Specifically, I thank Dr. Liqiang Nie, Geng Xue, and Xuemeng Song for their invaluable discussion during our lab meetings. I would also like to thank Alexander Farseev for all the discussion we had during our PhD studies.

Contents

Abstract

The increasing popularity of social media has encouraged health consumers to share, explore, and validate health and wellness information on social networks, which provide a rich repository of Patient Generated Wellness Data (PGWD). While data-driven healthcare has attracted a lot of attention from academia and industry for improving care delivery through personalized healthcare, limited research has been done on harvesting and utilizing PGWD available on social networks. This thesis focuses on learning wellness profiles of users, both at micro-level of individuals and macro-level of communities. Towards this end, we propose a unified framework and algorithms to perform the following tasks.

(1) To extract the wellness information of users, we propose a learning framework that utilizes the content information of microblogging messages as well as the relations among event categories to categorize messages into a wellness taxonomy.

(2) To learn the latent profile of users, we propose an approach which directly learns the embedding from longitudinal data of users, instead of vector-based representation. In particular, the proposed framework simultaneously learns a low-dimensional latent space as well as the temporal evolution of users in the wellness space. To construct an effective framework, we incorporate two types of wellness prior knowledge: (a) temporal progression of wellness attributes; and (b) heterogeneity of wellness attributes in the patient population. The proposed approach scales well to large datasets using parallel stochastic gradient descent.

(3) To learn the profile of user groups, we first integrate different social views

of the network into a low-dimensional latent space representing users' profiles. We then learn the optimal community structure by imposing a similarity constraint over the affiliation vectors of the users, which seeks dense clusters of users in the latent space. We seamlessly incorporate prior knowledge about the community structure into the community discovery process and turn the process into an optimization problem, where community profile is constructed using a linear pooling operator integrating the profiles of the members.

To evaluate the effectiveness of the proposed framework, two large scale datasets were constructed by crawling social activities of diabetes patients in Twitter. Extensive experiments have demonstrated: (1) the importance of modeling both content information and events relation in wellness event extraction; (2) the significance of joint modeling temporality of wellness features and heterogenity of the user in wellness profiling; (3) the importance of fusing all social behaviors for community discovery and profiling.

List of Tables

List of Figures

CHAPTER 1

Introduction

The past decade has recorded a rapid development and change in the Web and Internet. We are currently witnessing an explosive growth in social networking services, where users are publishing and consuming online contents. In such a context, millions of users, every day, publish their posts in different online social networks (OSNs), such as Twitter, Facebook and Flickr. For example, today, more than 56 percent of American adults older than 65 years use social media, which records an increase of more than three times compared to 2010 when only 11 percent had been reported [1]. Users in social media enjoy a wide range of freedom. Thus, they can freely publish their opinion and easily connect to their friends. As a result, people constantly share and discuss about various topics from personal events like birthday party, to public event like Ebola outbreak, to daily events like going to office.

In such a context, health consumers increasingly utilize social platforms to

[1] http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/

fulfil their health demands through seeking and sharing health information and experiences as well as providing online social support for their peers (Attai et al., 2015, Davis, Anthony, and Pauls, 2015, De Choudhury, Morris, and White, 2014). For example, it was reported that 57% of e-patients with chronic conditions constantly and actively refer to social media to acquire health information while 20% of them have already participated in generation of online health contents [2]. The emerging of self-tracking gadgets and the enthusiasm of users in taking informed health decisions has also intensified this trend. This motivates users to disclose their health information in social platforms (De Choudhury et al., 2013). For example diabetic patients frequently post about their health conditions, medications, and the outcome of medications on social media platforms like Twitter and Instagram. Further, the ubiquity of social media encourages health consumers to not only discuss about their health conditions and share experiences but more importantly share their health related attributes and measurements, like blood pressure and blood glucose, which provides an invaluable resource to study and analysis individual's and communities' wellness and behaviors. Figure 1.1 depicts several examples of disclosing wellness information in Twitter, where people publish detailed measurements and values about their activities, food consumption, and their health attributes, e.g. blood glucose values. While Electronic Health Records (EHRs) are increasingly utilized in medical informatics as an important and distinct data source, limited research efforts have been devoted into utilizing

[2] http://www.pewinternet.org/2010/03/24/social-media-and-health/

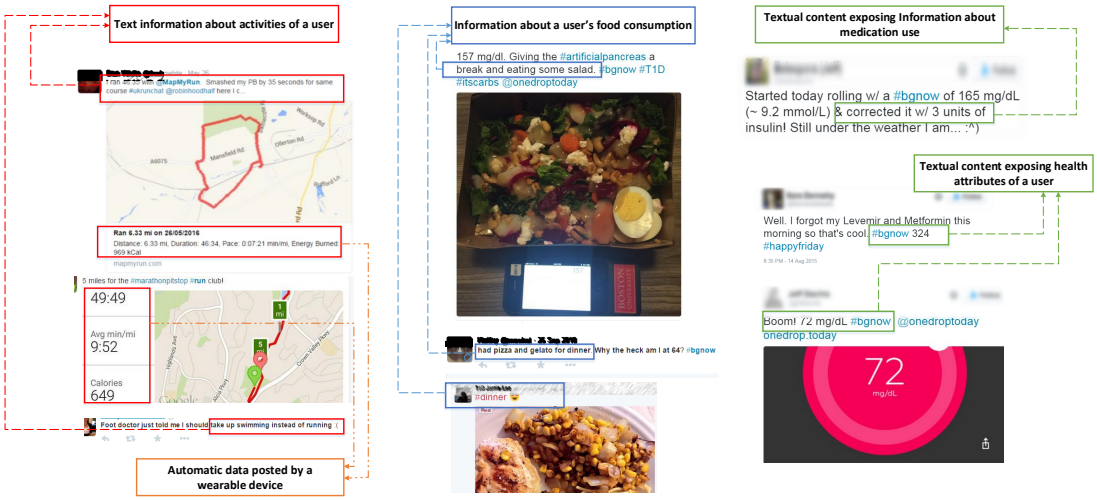


Figure 1.1: Sample of wellness information on Twitter; (left) information about activities; (middle) information about food consumption; (right) information about wellness of users. The advent of wearable devices has resulted into device-generated contents as shown in the left column.

Patient-Generated Wellness Data (PGWD) available on social networks (Che et al., 2015, Sun, Wang, and Hu, 2015, Robinson, 2012).

Concurrently, rates of chronic diseases, often referred to as non-communicable diseases (NCDs), continue to drastically rise worldwide. In 2001, chronic diseases contributed to approximately 60% of the 56 million deaths in the world [3] and the burden of the diseases is expected to increase 11% by 2020, alarming the needs for controlling its increase. In health sciences, there is an inevitable consensus that chronic diseases are largely preventable diseases by lifestyle intervention (Kuh and Shlomo, 2004). In essence, almost two-thirds of non-communicable diseases deaths are linked to unhealthy habits such as tobacco use, unhealthy diets, physical inactivity and harmful use of alcohol. Therefore, there is a high desire to provide computational systems which are able to assist people in managing their lifestyle

[3] http://www.cdc.gov/chronicdisease/overview/

and wellness. The emergence of several online services, such as "myfitnesspal" and "myfooddiary", and mobile application, such as "diet assistant" and "calorie counter", is indeed an effort to answer this demand by assisting people in understanding and improving their lifestyle and wellness.

The boom in PGWD opens up a vast range of possibilities to study individual's wellness and collective behaviours of social groups on an unprecedented scale. A major aspect of building an effective information and social service is the construction of user profiles that accurately represent users' interests and behaviors. User profiles are often used in personalization of services, outputs, and content to individuals needs and preferences towards gaining maximum user sanctification. As such, in this thesis, we focus on wellness profiling of users and communities, where we harvest social media to identify, understand, and predict, wellness attributes and behaviours of users and groups.

The remainder of this chapter is organized as follows. It first introduces the background of user profiling, in general and wellness domain, in Section 1.1. It then continues by highlighting motivations and illuminating challenges in Section 1.2, and Section 1.4, respectively. It continues by briefing the contributions in Section 1.5. It then ends by introducing the notation we use in Section 1.6, followed by the outline of the thesis in Section 1.7.

1.1 Background

In this section, we briefly explain the background context of the current research work. Indeed, user profiling has been largely utilized to boost the effectiveness of various information and social services. Briefly speaking, the related techniques and approaches can be divided into three categories: user profiling, group profiling, and vertical domain profiling.

1.1.1 User Profiling

User profile learning plays an increasingly important role in many social media applications. Learning a user profile can assist us in better modeling user's behaviours and interests which are essential in many social media applications such as recommendation services. Generally, a user profile is utilized to provide personalized services and results. Broadly speaking, user profile learning techniques can be divided into two broad categories: explicit and implicit profiling approaches.

Explicit user profiling focuses in inferring personal and social attributes and characteristics of users, such as gender, race, occupation, education, location, political affiliation, and so on. Most of existing studies focus on user-centric information, such as published tweets (Burger et al., 2011, Cheng, Caverlee, and Lee, 2010, Paul and Dredze, 2011), blog entries (Burger and Henderson, 2006, Yan and Yan, 2006, Nowson and Oberlander, 2006), browsing histories (Jones et al., 2007), and other types of user generated contents to profile users' certain attributes.

Among users attributes, age and gender (Peersman, Daelemans, and Van Vaerenbergh, 2011, Rao et al., 2010, Schwartz et al., 2013), occupation (Preoţiuc-Pietro, Lampos, and Aletras, 2015), and location prediction (Cheng, Caverlee, and Lee, 2010, Hecht et al., 2011) have attracted many researchers. Recently, network centric data has also been used for user attribute prediction, aiming to utilize user social friends, also referred as social dimensions, to profile users more effectively. Lately, multi-source user profiling has been investigated for effective user profiling, where users' data from multiple social networks are harvested for attribute profiling. For example, Farseeve et. al (Farseev et al., 2015) profile users' gender and age using a supervised classification approach based on users' published posts on multiple social networks, i.e., Twitter, Instagram, and Foursquare. The main challenge in multi-source user profiling is how to integrate heterogeneous information from various source into a unified learning framework. Overall, early fusion and late fusion are two major techniques for integrating multiple data sources (Akbari, Nie, and Chua, 2015).

Implicit user profiling, however, learns a latent representation, often a kind of distributed representation, for each user, which is capable of discriminating some aspects of user behaviours and interests. Matrix factorization (MF) (Koren, Bell, and Volinsky, 2009,Moghaddam, Jamali, and Ester, 2012), and topic models (Blei, Ng, and Jordan, 2003, Ding, Li, and Peng, 2008) are two main techniques which have been widely used for learning a latent representation of users and items in recommendation system. The hypothesis behind latent representation learning is

that users and items can be mapped into a low-dimensional space representing their relations. Recently, implicit feedback of users, such as user activities, is also utilized for inferring user preferences (Gupta and Singh, 2015).

1.1.2 Group Profiling

Group profiling aims at learning the collective behaviour of a group of people, also known as community in social media computing. It can be applied for policy-making, direct marketing, tracking interest shifts of communities, network visualization and navigation (Cruz, Bothorel, and Poulet, 2013,Omidvar-Tehrani, Amer-Yahia, and Termier, 2015). Group profiling in social media and data mining is often modeled as a two stage framework including community discovery phase and the aggregation of attributes of the community members. Due to its importance, several research efforts have been devoted into community discovery problem. For example, modularity decomposition has been applied to link information in social networks (Yang et al., 2009), and communication pattern has been used for expert team detection (Lappas, Liu, and Terzi, 2009), and generative models were used to find topical communities (Zhou, Jin, and Liu, 2012). Community profile is then constructed by aggregation of shared group interests. For example, Wang et. al (Wang, Guo, and Lan, 2014), have aggregated the meta data of users and venues to characterize the communities based on their members' location interest. Similarly, Zhao. et. al, integrated venues, images, and comments to learn a multi-modal community profile in location-based social networks (Zhao et al.,

2013b).

While two stage group profiling decomposes the problem into two intuitive sub-problems, how to infer group profiles from discovered communities pose a great challenge and still is an open problem.

1.1.3 Vertical Domain Profiling

User profiling has also been leveraged in vertical domains such as apparel domain and shopping domain (Geng et al., 2014, McAuley et al., 2015, McAuley, Pandey, and Leskovec, 2015). As we are interested in the vertical domain of wellness and health of users, we only discuss related concepts of wellness profiling in this section. Social media platforms have proven their importance as a convenient tool for broadcasting information, sharing opinion and thought, and interacting with friends. The success of social media have attracted the researchers from health and medical community, in particular public health, to study the health of individuals and of population, aiming at developing services and policies that improve the wellness of users (De Choudhury et al., 2013, De Choudhury, Counts, and Horvitz, 2013b). Social media data has been also utilized for profiling users and communities from wellness aspect. In the individual aspect, user generated contents and online behaviors have been utilized for profiling user wellness attributes such as depression (De Choudhury et al., 2013), stress (Lin et al., 2014), post-partum behavioral changes (De Choudhury, Counts, and Horvitz, 2013b), shifts to suicidal ideation (De Choudhury et al., 2016), etc. In population-level, so-

cial media data has been employed as a useful resource for studying population behaviour related to diabetes, obesity, cardiovascular disease (Eichstaedt et al., 2015), just to name a few. Most existing research efforts have used social media data for analyzing wellness of population in country or city scale. For example, in (Paul and Dredze, 2011), researchers predicted the rates of diabetes and obesity for 15 US cities. Similarly, Foursquare has been examined for the group obesity prediction and cultural characterization (Abbar, Mejova, and Weber, 2015). We will discuss the related literature with more details in chapter 2.

1.2 Motivation

Learning the wellness profile of users can assist individuals and communities to improve their wellness and lifestyle. At individual level, it can provide better online services which assist users in different ways. First, people increasingly utilize social media for healthcare; they share their daily activities and health information in social platforms. The wellness profile of a user can be used for several personalized online services assisting them in making informed health decision. For instance, the wellness profile of a user can be used in a health social network for recommending contents related to their health conditions. Taking diabetes as a wellness problem, the system can recommend a Type II diabetes patient with content related to his specific condition. Further, in case of having a question, the system can route the question to a proper expert or a patient with similar condition to obtain first hand suggestions and experiences. Second, considering the

popularity of wearable devices, the wellness profile of a user from social networks can complement his profile extracted from the wearable device. Indeed, collecting user information from multiple sources provides a comprehensive understanding of user's behaviors and interests, which improves the efficacy of the system.

At group level, it provides insights about population wellness. In particular, an important attribute of social media data is that posting are performed in a naturalistic way in the course of daily activities and events. By collecting and aggregating wellness information of users, discovering potential communities, and profiling the discovered communities, we can study the wellness of communities which provides insights about the wellness and health of the population. It can be utilized for policy-making, trend analysis, and search and tracking wellness groups. This data complements the information and insight we can obtain through traditional methods like completing surveys by a group of users as a population sample.

1.3 Problem Definition

Given social media accounts of users, the problem we aim in this thesis is to make sense the wellness of users on online social networks. In particular, we investigate approaches to learn the wellness profile of users, where we leverage social media data to identify, understand, and predict wellness attributes and states of users and communities. To achieve this end, we first propose an approach to extract posts which directly indicate the happening of personal wellness events for the users. We

then study how social content and behavior of users reflect their wellness attributes and states. Following this line of research, we propose to learn the wellness profile of individuals from their social media contents and behaviors, identify the user communities discussing about wellness on social media platforms and build the wellness profile of user communities.

1.4 Challenges

Due to the importance and value of wellness profiling for service providers, it has attracted a lot of research interest in social media computing and health informatics. However, learning wellness profile of users in social media is a nontrivial task due to the following reasons.

Limited data. A major challenge for a vertical user profiling approach is the *limited data* problem. Looking from macro-level, social media data is known as a big data since millions of users are continually generating content and interact with each other. However, from a micro-level perspective, many social media users do not generate sufficient data or they partially active in content generation. Considering the specific domain of wellness, most of the time they are not keen enough to participate in content generation or they hide their wellness data due to privacy concerns.

Noisy and Imbalanced Data. The language used in social media is highly varied, informal and full of slang words, which makes the problem more chal-

lenging as compared to standard text processing problems like document clustering. Meanwhile, users frequently post their personal significant events together with trivialities and other public events, which makes the wellness related posts relatively sparse and rare in social media contents. Identifying wellness information from a huge volume of other non-health related events is a big challenge. Even though we can successfully identify the personal health events, it is still time and labour consuming to label them.

Heterogeneity. Social media data is heterogeneous meaning that various kinds of information are available in social networks such as text, image, and video contents. Integrating different kinds of information in the learning model is essential and challenging. Further, wellness domain is characterized by the heterogeneity of patient population which means that wellness attributes and events related to each user can be highly different from the others according to their demographic attributes (e.g., age and gender), type of disease (e.g., Type I Diabetes, Type II Diabetes, Gestational Diabetes, etc.), and many other behavioural and genetic factors. Although proper modeling of heterogeneity in contents and users is essential, it is still an open problem in machine learning and data mining studies.

Longitudinality. Wellness data are longitudinal per se, which means multiple measurements or repeated events are available for each subject. For example, Hemoglobin A1c (HbA1c) test might be done several times per year for each diabetic patient. The current measurement of wellness attribute along with

the patient history both are important to construct a precise user model. To effectively utilize this data, temporal dependency of the features needs to be appropriately modeled in the learning framework.

1.5 Contributions

The contributions of the thesis can be summarized as follows.

1. We propose a framework which extracts personal wellness events from published posts of users in microblogging social services. The proposed framework utilizes content information of published posts as well as the relations between different wellness categories to categorize messages into a wellness taxonomy.

2. We focus on diabetes as a specific type of wellness problem and perform deep analysis of users online behavior and characteristics, which provides better insights about capability of social media platforms in understanding and profiling user's wellness. In particular, we study the the behavioral distinction between diabetic patients in order to characterize those who can successfully manage their chronic conditions and those who fail. We have observed several distinctions in terms of linguistic, textual, and visual contents of published posts online. Based on the findings, we propose two different intervention method for assisting users better adopt their chronic condition.

3. We propose an approach to automatically learn the embedding of users di-

rectly from their longitudinal data. The proposed method takes into account two types of wellness prior knowledge: temporal progression of wellness attributes; and heterogeneity of wellness attributes in the patient population. Taking diabetes as an example of wellness domain, we conduct extensive experiments to evaluate our framework at tackling three major tasks in wellness domain: attribute prediction, success prediction and community detection.

4. We proposed a community discovery and profiling approach for social media users. The proposed model simultaneously learns the profiles of users and their affiliation to communities in a low-dimensional space, which is constructed from the integration of different social views of the network. Extensive experiments on a real-world dataset of diabetic patients have demonstrated the effectiveness of the proposed approach on discovery and profiling communities as well as leveraged several interesting insights about users' interactions in social media platforms.

1.6 Notation

In this dissertation, we use the following notations to describe the formulation of the problems and provide mathematical models. We use boldface uppercase letters (e.g., $\mathbf{A}$) to denote matrices, boldface lowercase letters (e.g., $\mathbf{a}$) to denote vectors, and lowercase letters (e.g., a) to denote scalars. The entry at the i-th row and j-th column of a matrix $\mathbf{A}$ is denoted as $\mathbf{A}_{(ij)}$. $\mathbf{A}_{(i*)}$ and $\mathbf{A}_{(*j)}$ denote the

i-th row and j-th column of a matrix $\mathbf{A}$, respectively. Meanwhile, subscript, i.e. $\mathbf{A}_i$, is used to denote the i-th item in a set of items $\mathcal{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_i, \ldots\}$. $\|\mathbf{A}\|_1$ is the ℓ_1-norm and $\|\mathbf{A}\|_F^2$ is the Frobenius norm of matrix $\mathbf{A}$. Specifically, $\|\mathbf{A}\|_1 = \sum_{i=1}^{m} \|\mathbf{A}_{i*}\|_1$ and $\|\mathbf{A}\|_F^2 = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |\mathbf{A}_{ij}|^2}$. Besides, $tr(.)$ denotes the trace of a matrix.

1.7 Thesis Structure

The remainder of the thesis is organized as follows.

Chapter 2 describes work related to user wellness profiling. It first reviews recent research efforts on event detection from social media platforms in two major categories: public event detection and personal event detection. It next introduces related work in user profiling followed by introducing major advancements in group profiling.

Chapter 3 demonstrates an approach for mining wellness information of users from social media platforms. It first formally define the problem of personal wellness event extraction from social media and then provides a supervised approach for extracting personal wellness events which are mentioned in published posts of users, aiming at constructing a personal timeline of wellness events for each social media user.

While Chapter 3 demonstrates the capability of social media platforms as a social sensor for inferring users' lifestyle and wellness, in general, Chapter 4 focuses on a specific kind of health condition, diabetes, to deeply analyze the

capability of social media in wellness profiling and gain better insights about this connection. In particular, Chapter 4 presents a characterization study on the behavioral distinction of two groups of diabetes patients, aiming at distinguishing diabetes patients who can control their blood glucose level and those who fail to manage their blood glucose level, most of the times.

Chapter 5 demonstrates the importance of feature learning from longitudinal data. It proposes an approach which directly learn the representation of users in the latent space, where both temporality of wellness attributes and heterogeneity of patient population are jointly modeled.

Chapter 6 introduces a learning framework which simultaneously learns the profile of communities in social networks. The framework leverages prior knowledge and behavioral factorization to learn a low-dimensional latent space from online behavior of users. Community discovery is then performed in the latent space by considering all social behaviors of users.

Chapter 7 concludes the thesis with highlights of contributions and offers pointers for potential future research.

CHAPTER 2

Background and Literature Review

Learning Profile of users, in micro-level of individual and macro-level of groups, are related to a range of research directions. This chapter serves to introduce major retrospective studies related to this dissertation. To do so, we first review recent works on event detection in Twitter social platform in Section 2.1. We next introduce related work in user profiling in Section 2.2, followed by Section 2.3 which reviews related efforts in group profiling.

2.1 Event Extraction in Social Media Platforms

Efforts on event extraction and timeline generation can be roughly categorized into two broad classes: public event extraction and personal event extraction. In essence, most existing studies on event detection in social media platforms focus on detecting public events due to the initial belief that social platforms are designed for sharing opinion and spreading information. However, personal event detec-

tion has recently attracted attention in personalized information services. In the following subsections, we briefly describe major advancements in both directions.

2.1.1 Public Event Detection

Retrospective studies in social media computing have made it evident that social media streams can provide vital information related to events in real-time. Hence, event detection from online social networks has received considerable interest in the fields of data mining and knowledge discovery (Petrović, Osborne, and Lavrenko, 2010, Phuvipadawat and Murata, 2010, Diao et al., 2012). Twitter, amongst other social platforms, has attracted higher attention due to the rich user-generated text data that can be accessed in real-time (Petrović, Osborne, and Lavrenko, 2010). Depending on the task, several approaches have been proposed to detect events by exploiting different aspects of the data such as content, temporal and social features. Broadly speaking, existing public event detection algorithms can be classified into two major categories: document-pivot methods and feature-pivot methods. Document-pivot techniques cluster the documents into several categories based on their semantic distances while term-pivot methods rely on the distributions of words and discover events by detecting burstiness in word groups.

In this domain, Petrovic et al. (2010) proposed an adaptation of document-pivot approach to first story detection in fast social streams like Twitter. Their framework utilizes the locality sensitive hashing (LSH) with the cosine similarity to compute the similarity of the incoming tweet to existing tweets. Similar tweets

are then grouped into a cluster demonstrating an event. Similarly, Phuvipadawat and Murata (2010) proposed an approach for detecting and tracking breaking news from Twitter. Their proposed approach first samples tweets which contain specific hashtags and keywords, such as "#breakingnews", and then clusters similar tweets together to form a news story. To effectively compute the similarity between short messages, they utilized a variant of Tf-Idf boosted by a Named Entity Recognizer (NER).

Yet another way to cluster tweets into events is to employ topic modeling techniques to extract hidden topics from streams (Blei, Ng, and Jordan, 2003, Ramage et al., 2009). As standard topic modeling approaches do not consider temporal information, several studies have been conducted to incorporate time dimension in topic models. For instance, Diao et al. (2012) proposed a new topic model that integrates both the temporal information of microblog posts and users' personal interests. Bursty topic was then identified as a set of tweets that contain similar words, i.e., discussing on same topic, however they published by different users. While topic models provide a principled way to detect hidden topics in a text collection, most exsisiting approaches can be applied in a retrospective and offline manner. Further, they depends on parameter stetting and large number of sampling iterations.

TwitterMonitor (Mathioudakis and Koudas, 2010) treats bursty keywords, that suddenly appear in tweets at an unusually high rate, as indicators of a new trending topic. It first identifies bursty keywords and groups them based on their co-

occurrences to get trends, i.e., keyword groups. For each trend, singular value decomposition (SVD) and entity extraction techniques are then utilized to build a better contextual description of the trend.

Graph-based approaches have also been used for clustering keywords pertaining to their pairwise similarities. For example, Sayyadi, Hurst, and Maykov (2009) proposed to build a graph named KeyGraph based on keywords co-occurrence in documents. In their method, keyword clusters are then discovered using a community detection algorithm based on betweenness centrality. Graph-based approaches have also been applied to the task of tag clustering in the context of collaborative tagging systems aiming at discovering groups of tags pertaining to topics of social interest (Papadopoulos, Kompatsiaris, and Vakali, 2010).

Alternatively, Weng et al. (Weng and Lee, 2011) captured the burstiness of words by considering them as signals and applying wavelet analysis on consecutive time slots. There were three steps in their approach. First, they applied wavelet transformation and autocorrelation to measure the bursty energy of each word, where the words with low energies are filtered. Second, cross correlation was applied to measure the similarities between event features. Finally, modularity-based graph partitioning was utilized to detect the events with high cross correlation amongst words.

Recently, Li, Sun, and Datta (2012) argued that multi-word segments or word n-grams, as compared to single words, can improve the performance of the event detection task. They proposed to split each tweet into non-overlapping segments,

i.e. phrases possibly refereeing to named entities or semantically meaningful units. The bursty segments were ranked and clustered into candidate events. Finally, Wikipedia was exploited to identify the realistic events and to derive the most newsworthy segments to describe the identified events.

2.1.2 Personal Event Detection

As compared to public event extraction, the work on personal event extraction from social media contents is still relatively sparse. Most existing approaches focus on extracting personal events in order to construct a personal timeline of the event.

Prior research efforts mainly focused on clustering and sorting information of a specific person from web search (Al-Kamha and Embley, 2004,Kimura et al., 2007, Wan et al., 2005,Yoshida et al., 2010). To detect personal events, Diao et al. (2012) designed an extension of Latent Dirichlet Allocation (LDA) that separates personal topics from public burstly topics. It considers both the temporal information of microblog posts and users' personal interests. Recently, Li and Cardie (2014) utilized multi-level dirichlet process model to construct a chronological timeline for individuals based on their published tweets. Their system does not recognize the categories of event however identifies tweets that are personal and time-specific. Besides if the topic does not adequately discussed by their local network, it cannot be detected since topic models use word frequency to detect topics.

TwiCal is a system which extracts events with their description from Twitter

posts of users (Ritter et al., 2012). Each event in TwiCal is represented by 4 attributes includes: named entity, event phrase, event type, and calender date. Given a raw stream of tweets, the system classifies named entities, and event phrases into event types. After detecting temporal expression in tweets and resolving them, events are scored based on association between each entity and the specific data.

Recently, a supervised classification approach was proposed in (Li et al., 2014), which detects major life events from tweets. Their framework exploits congratulations and condolences speech acts to extract training examples. Their system is comprised of three components. First, it identifies the major categories of life events and filters out mundane and irrelevant tweets. Second, the system distinguishes personal events from the events that involve other users, i.e., extract personal events. Finally, the properties and attributes of events such as location is identified as a description for the event. The major drawback of their framework is that it combines several components sequentially which results in the propagation of errors from each component to the next one.

In this line of of research, several supervised approach have been proposed to examine different features for automatic detection of life events from short messages (Choudhury and Alani, 2014, Dickinson et al., 2015, Cavalin, Moyano, and Miranda, 2015). Lately, Liu et al. (2015) proposed a model to extract patient experience from health forums, where they classified each sentence as containing patient experience or not containing patient experience. Their system utilizes

the global context and local context of each post to extract patient experiences. However, their approach is not applicable to tweets due to the short length of messages.

Seeing from developed applications, the increasing popularity of social media leads to high demands on services for tracking individual such as the famous Facebook timeline. To fulfill this demand, Facebook has recently generated 270 million of 1 minutes look back videos from users' timelines and over 200 million users watched their look back movie in the first two days[1]. Similar projects like Intel's Museum of Me[2] attempted to follow a similar direction by collecting UGCs in social platforms.

To summarize, past efforts generally focus on discriminating public events and personal events based on word search or topic models. Therefore, they only detects major life events such as graduation and marriage. However all lifestyle event are not major life event and they are not discussed in local circles of users.

2.2 User Profiling

Modelling users' behaviour and identifying their interest is an important aspect of constructing an effective recommender and information system, which is often referred as "User Profiling". This is a crucial requirement in improving the performance of system as well as the satisfaction of users by providing personalized

[1] `https://code.facebook.com/posts/236248456565933/looking-back-on-look-back-videos/`
[2] `http://www.intel.com/museumofme/en_AU/r/index.htm`

recommendations and services. For example, profiling a user's location or topic of interest permits us to provide personalized search results in search engines, local news in news sites, and targeted ads in advertisement systems. As previously discussed, user profiling techniques can be divided into two broad categories: explicit and implicit profiling approaches. The former attempts to detect different attributes of a user, such as age and gender, based on the user data, while the latter models the user's interests and behaviour in order to provide personalized results.

2.2.1 Explicit User Profiling

Based on the type of information used in profiling, literature can be divided into following categories: *content-based profiling*, *network-based profiling*, *hybrid-base profiling*, *co-profiling*, and *multi-source user profiling*.

Content-based User Profiling

The traditional approach in user profiling is to utilize user-generated contents, such as status updates, user profiles and shared multimedia contents, to infer user attributes and preferences. Usually, the problem is modelled as a supervised learning task and various classification approaches are used to predict users' attributes. Support vector machines (SVMs), probabilistic modelling, and boosted decision trees are the most common algorithms which are used for solving the problem.

Early in this direction, Hecht et al. (2011) conducted an investigation to verify

the information embedded in the location field of users' profiles in Twitter. They reported that many users provide no information or non-real information in their profiles and demonstrated that the explicit location sharing behaviors should be examined in the context of implicit behaviors. They proposed a Multinomial Naive Bayes model to classify tweets and attempted to predict the home country and state of Twitter users. Their experiments used a limited dataset of 4 countries, and the state-level experiments were restricted to the United States. They did not make use of geotags but instead obtained location data from the users' profile location fields. After filtering out users with fewer than 10 tweets, a dataset of almost $100,000$ users remained. Their approach correctly placed users in their home states with an accuracy of up to 30%, and in their home country with an accuracy of up to 80%. The model correctly placed the users at a much better accuracy than random, indicating that users implicitly reveal location information in their tweets.

Similarly, Cheng, Caverlee, and Lee (2010) proposed a probabilistic framework for predicting a Twitter user's city-level location merely based on tweet content. The proposed framework applied two basic improvements:

- feature selection was utilized to automatically identify words in tweets with a strong local geo-scope.
- A lattice-based geographic smoothing method was used to refine a user's location estimation.

To evaluate the framework, city models were built from over 4 million tweets

from approximately 131,000 users who had a declared location in a city in the United States. They reported an accuracy of 51% where users correctly placed within 100 miles of their correct location. Further, Bo, Cook, and Baldwin (2012) proposed feature selection for finding location indicative words (LIWs). They demonstrated that finding LIWs boosts the performance of text-based geolocation task, outperforming state-of-the-art geolocation prediction methods by 10.6% in accuracy and reducing the mean and median of prediction error distance by 45 km and 209 km on a public dataset, respectively.

Network-based User Profiling

Network-based approaches utilize users' social graph, i.e., friendship, or interaction information to estimate their preferences. The hypothesis behind network-based user profiling springs from the simple but effective social theories of *heomophily* and *contagious*, which attests that interests of people are correlated with that of their social connections and friends. Relational learning or collective classification are widely used to refer to the task of classification in the network data when data are linked. Indeed, the major characteristic of network data is that independently identical distribution ($i.i.d$) assumption is invalid and data instances are dependent to each other. Hence, it is possible to capture the correlation among data instances to improve the performance of learning framework. Normally, a relational classifier is constructed, based on the correlation of the features and labels of the data, to minimize the discrepancy between labels of connected data

instances.

Mislove et al. (2010) demonstrated that the missing attributes of users can be inferred from the attributes of other users in an online social network. They found that users with common attributes often form dense communities, which motivated them to propose a method of inferring user attributes by detecting communities in social networks. Using Facebook data, they demonstrated that when only 20% of the users in the network provide their attributes, it is possible to estimate the attributes for the rest of the users with an accuracy of over 80%.

Backstrom, Sun, and Marlow (2010) studied the relationship between proximity and freindship on social media platforms and observed that, as expected, the likelihood of friendship drops monotonically as a function of distance. Upon this observation, they proposed an approach for predicting the physical location of a user, given the known location of his friends. They examined the probability of friendship as a function of distance and found that:

- The probability decreases as distance between users increases.
- The probabilities fit an exponential distribution.

They used a dataset of Facebook users including 2.9 million users whose locations were known. Using a maximum likelihood approach and a simple assumption of close proximity of friends, they predicted the physical location of 69.1% of the users with 16 or more located friends to within 25 miles.

Hybrid-based User Profiling

Recently hybrid-based profiling attracted much interest in user profiling. The hypothesis behind hybrid-based user profiling is that we can integrate both content and network information for better user profiling.

Li et al. (2012a) proposed a unified discriminative influence model, named as UDI, to predict the home location of users in the context of Twitter social network. Their framework integrates user's generated contents (tweet posts) and users' social networks (friends) in a directed heterogeneous graph to overcome scarce and noisy data problem. They developed two variants of their framework named local prediction model and global prediction model. The former integrates locations based on a user's friends, followers and tweets in a discriminative way to estimate the location. The latter extends the local method with additional unlabelled data, i.e., in this case unlabelled users. Experimental results on a large scale dataset demonstrated the ability of the approach by 13% improvement on sate-of-the-art baselines. Similarly a generative model, named multiple location profiling model (MLP), has been proposed to profile location of users on Twitter (Li, Wang, and Chang, 2012). In particular, they utilized two probabilistic models to predict locations with contents and friendships. As a user may related to multiple location, they introduced two mixture models to capture the noisy information. Experiments on a large-scale dataset demonstrated 10% improvement in accuracy of estimating user's home location.

Co-Profiling

Co-profiling integrates user profiling with other related tasks, such as relation prediction, to achieve mutual enhancements. In this line of research, Dalvi, Kumar, and Pang (2012) investigated the problem of object matching for tweets, where the aim is to generate a list of tweets corresponding to an objects. In (Dalvi, Kumar, and Pang, 2012), a probabilistic model was proposed to infer the matches between tweets and restaurants from a given list. The authors assumed that : (1) each user is interested about a set of objects; (2) each user and object are associated with a geographic location; and (3) the probability that a user tweets about an objects depends on: (a) user's interest on the object; (b) the popularity of the objects, and more importantly (c) the inverse of the distance between the user and the object locations. Based on these assumptions, they proposed a probabilistic model consisting of two components: a distant model and a language model. Experimental results demonstrated the gain of their model in compared to non-geographical model in inferring the location of the users accurately.

In another work, co-profiling of attributes and relations was investigated (Li, Wang, and Chang, 2014). Briefly, the authors captured the correlation between attributes and relation types. Their framework was based on two intuitive assumptions as follows:

- Attribute Profiling. Different types of relationships propagate different attributes. For instance, university should propagate from college mates, while

occupation should be propagated from colleagues.

- Relation Type Profiling. Relationship types between users can be identified by their shared attributes and network structure. For example, a set of friends who are strongly connected and share occupation might be colleagues.

Upon these assumptions, they proposed a label propagation approach to infer the missing attributes and relations based on users' known labels. To do so, a cost function was designed to model both types of dependencies and alternative optimization has been leveraged to solve the minimization problem (i.e. minimizing the cost function). Their results demonstrated that co-profiling algorithm not only profiles users' attributes accurately, improving the state-of-the-art methods greatly, but also correctly identifies latent circles of users' friends, which are useful for many advanced applications.

Multi-source User Profiling

In recent years, multi-source profiling has attracted attention for learning profiles of users from several online social networks. For example in earlier work, Yuan et al. (2012) studied the approaches to integrate multi-modal data from clinical measurements to enhance the results of Alzheimer's Disease prediction. However, the model was formulated as a binary classification tasks, and, thus, not directly applicable to real-world scenarios. Later, Song et al. (2015a) proposed a learning framework for prediction of users' Volunteer tendency from multiple social networks. They extracted information cues about the users form social networks

such as, demographic information, practical behaviors, historical posts, and profiles of social connections. As the social network data are often incomplete, they mitigated the problem arising from missing data through inferring missing data by learning a latent shared space using constrained Non-Negative Matrix Factorization (NMF). Authors finally predicted the volunteerism tendency of users using a regression model fusing all the data from multiple sources. The learning framwork models the problem as a binary classification problem which is the main drawback of the proposed learning framework. Later, Song et al. (2015b) proposed a multi-source multi-task framework which infers users' interests. The proposed framework exploited the prior knowledge about users' interests in a tree-structure which was used as a regularizer in the proposed optimization problem.

At the same time, (Farseev et al., 2015) introduced efficient ensemble learning solutions, aiming to combine multi-source multi-modal data for demographic and mobility user profiling, respectively. The above model performed the best among various state-of-the-art baselines and demonstrated the necessity of learning from multiple sources to improve the user profiling performance. Lately, deep learning was also used for fusing multiple social network data to predict users' volunteerism tendency.

2.2.2 Implicit User Profiling

Representation learning, or latent feature learning, is a popular approach for discovering low-dimensional structure from high dimensional data. We are interested

in factorization based models which aim at finding a low rank decomposition of original space approximately recovering the original space including sparse coding, Singular Value Decomposition (SVD), Principal Component Analysis (PCA), Weighted Matrix Factorization (WMF), and so on (Chen et al., 2015, Lin et al., 2009, Volkovs and Yu, 2015). Recently, latent factor decomposition has been attracting much interest to alleviate data sparsity in recommendation task where user-item matrix is used to model user interests and intentions (Cai et al., 2011,He et al., 2014b, Jin et al., 2015, Zhao, McAuley, and King, 2014, Zhao, McAuley, and King, 2015). For example, Cai et al. (Cai et al., 2011) proposed a graph regularized NMF (GNMF) approach which employs the geometrical information of data space in factorization process. Similarly, semi-supervised GNMF (SGNMF) incorporates label information into the graph construction (Liu et al., 2012). NMF has also been applied onto multi-view data, where a shared latent factor was inferred from different views (He et al., 2014b, Pan et al., 2014, Song et al., 2015a). For instance, joint NMF has been applied to multi-view clustering of Web 2.0 items by decoupling the learnt latent factors inferred from different views (He et al., 2014b).

Personalization of latent factor modelling was first explored in (Shen and Jin, 2012), where a joint personal and social latent factor (PSLF) has been utilized for social recommendation. Similarly, Pan and Chen (2013) aggregated the features of a group of related users to reduce the uncertainty of the selected training instances. Zhao, McAuley, and King (2014) leveraged social connections to improve

the performance of one-class recommendation. Lately, they proposed a personalized feature projection method that employs users' projection matrices and items' factors to solve one-class recommendation problem (Zhao, McAuley, and King, 2015).

Most of the existing approaches for latent factor learning have been designed for vector-based representation to embed users (or items) in a low dimensional space. They will fail to provide effective representation if applied to longitudinal wellness data. Furthermore, existing feature learning approaches often assume that data items are *i.i.d.*, which is clearly violated in longitudinal data. Moreover, most of these approaches fail to model heterogeneity in data space or model temporal dependency as a regularized multi-task learning framework but overlook heterogeneity in data space. Our aim, in this research, is to learn a latent representation directly from longitudinal data where temporality and heterogeneity of data are jointly modeled.

2.3 Group Profiling

Group profiling aims at learning the collective behavior of a group of users that are also known as a user communitiy in social media computing. An intuitive approach for profiling at group level is to discover communities of users and then apply profile learning approach to capture information and behaviors of the community members (Zhao et al., 2013b). In other words, community profiling is often modeled as a two stage framework including community discovery problem and

aggregation of members' profiles to build the group profile. Group profiling is important for various applications including facility planning, recommendation (Zhao et al., 2013b), marketing (Qu and Zhang, 2013), and advertisement. In this section, we first overview the major advancements in community detection and then describe related works in learning the profile of groups.

2.3.1 User Community Detection

As mentioned above, it is important to find representative user communities in order to leverage on a group knowledge for various applications. Such task is commonly approached by modeling users' relationship as a graph, so that dense subgraphs of such graph can be treated as user communities. The graph can be constructed in several ways: (a) based upon *social connections between users* (i.e. follower/followie relationship) that are often hidden behind the privacy settings; (b) based on *user generated content*, when the edges of the graph are weighted as a distance between latent representations of users (i.e. cosine or Euclidian distance); or (c) as a *combination of the above two methods.*

After the construction of graph G, it is important to define: "what exactly a user community means". Traditionally, it has been modeled as a MinCut problem (Von Luxburg, 2007), where for a given number k of subsets, the MinCut, essentially chooses a partition $C_1, ..., C_k$ such that it minimizes the following ex-

pression,

$$cut(C_1, ..., C_k) = \sum_{i=1}^{k} W(C_i, \overline{C}_i) \tag{2.1}$$

where $\overline{C}_i$ stands for a compliment of C_i, and $W(A, B) = \sum_{u \in A, v \in B} \rho(u, v)$, and ρ is a distance function. Such a formulation allows us to find k user communities $C_1, ..., C_k$, that are grouped according to some criteria (defined by distance function ρ). The two most commonly used definitions of the MinCut problem are RatioCut and the so-called NCut (Von Luxburg, 2007) reformulation. The idea behind RatioCut is based upon an assumption that the resulting communities could have similar size, while NCut formulation constraints the sum of edges' weights in each community to be minimized among all communities. Both RatioCut and NCut are NP-Hard (Von Luxburg, 2007).

Fortunately, many approximate solutions exist for the problem above. One of the existing MinCut approximations, is the Spectral Clustering approach, which is defined as a standard trace minimization problem as follows,

$$\min_{U \in R^{n \times k}} tr(U^T \mathcal{L} U), s.t. U^T U = I. \tag{2.2}$$

According to the Rayleigh-Ritz theorem, the spectral clustering optimization problem can be solved as the first k eigenvectors of the normalized graph Laplacian $\mathcal{L} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where W is the adjacency matrix, and D is the degree matrix of the graph G. It should be noted that in cases that data comes from multiple

sources, the multi-layer generalization of Spectral Clustering and its variations can be used: $\min_{U \in \mathbb{R}^{n \times k}} \sum_{i=1}^{M} tr(U^T \mathcal{L}_i U), s.t. U^T U = I.$

In addition to the aforementioned techniques, other solutions of multi-source clustering problem can be utilized. For example, in (Farseev et al., 2014), the authors compared different conventional clustering techniques to solve the cross-source venue category recommendation task. Specifically, they detected user communities based upon data from Twitter, and used the obtained communities to perform cross-domain recommendation of Foursquare venue categories. The technique allows to tackle the Cold Start Problem in recommender systems, which allows to perform recommendation for users who fail to perform Foursquare activity. At the same time, Zhao et al. (Zhao et al., 2013b) proposed an approach to perform multi-modal venue recommendation based on regularized Modularity Maximization clustering. Following, same authors proposed another solution by imposing a regularizer on Matrix Factorization (Zhao, McAuley, and King, 2015).

2.3.2 Group Profile Inference

The aim of group profiling is to construct a descriptive profile illustrating the collective behavior of the members of the group, i.e., how the detected group looks like. Depending on the focused task, various approaches have been proposed for buliding a profile for a given group. In (Tang, Wang, and Liu, 2011), group profiling was formally defined as a set of attributes which can discriminate the group's members from the rest of users. The authors discussed on two major

properties of a good community profile as,

- *Descriptive.* The group attributes should reflect the foundation of the group illustrating the collective behaviors, shared interest, and the associated affiliation of members.

- *Robust.* Big data is characterized by velocity as huge stake of data is produced everyday. The arriving data tends to be noisy and so the profiling approach should be robust to the noise.

Following the above mentioned criteria, authors proposed three different approaches for profiling a group of users named: (1) Aggregation Group Profiling (AGP); (2) Differentiation Group profiling (DGP); and (3) Egocentric Differentiation Group Profiling (EDGP). AGP is the natural strategy which aggregates individuals' attributes. In other words, AGP seeks for attributes which are shared most frequently. They found that AGP fails to find an appropriate profile in noisy situations. To mitigate noise effect, they proposed DGP which differentiate a group form the rest of network to construct the group profile. EDGP is a variant of DGP with an egocentric view, which differentiates a group from their first hop neighbours in the network. Both DGP and EDGP achieved better performance in constructing group profile in noisy environment as compared to AGP. Although their approach is intuitive and effective, it demands explicit features and information and its performance was not evaluated in constructing latent profiles.

In another work, Wang et al. (2014) proposed an approach for overlapping community detection in Location-Based Social Networks (LBSN). Their frame-

work exploited both user-venue checkins information on the network and the attributes of users and venues to discover user communities. They adopted an edge-centric algorithms that utilized both inter-mode and intra-mode feature for clustering. After discovering communities, community profiling was performed by aggregating the metadata of users and venues that fall into the community. They first calculated the importance of each venue and user and then constructed the community profile by holding important users and venues. Meantime, they reported several interesting findings obtained through community profiling for three large cities, i.e., London, New York, Los Angeles.

Similarly, Zhao et al. (2013b) investigated to discover profilable communities from LBSN. They proposed a multimodal hypergraph learning approach to discover and profile social communities in LBSNs. In their framework, users, venues and posts, both textual comments and images, were considered as vertices of the hypergraph, while the related users, venues, text comments, and photos were connected by edges. Communities were detected by an efficient algorithm which detected dense subgraphs and profiles were constructed by weighted ranking of entities in each community. The approach was evaluated, both quantitatively and qualitatively, on a Foursquare dataset from Singapore and New York cities.

CHAPTER 3

Mining Personal Wellness Events from Social Media Platforms

Recent years have witnessed the revolutionary changes brought by the development of social media services through which individuals extensively share information, express ideas, and construct social communities. These changes can advance many disciplines and industries, and health is no exception (Nie et al., 2015, Lee et al., 2014). In such a context, many users are keen to share their wellness information on social platforms such as Twitter and Facebook (Hawn, 2009, Yang et al., 2014a, Dos Reis and Culotta, 2015, Paul et al., 2015). Take diabetes as an example; diabetic patients not only share about events happening around them but also frequently post about their current health conditions, medication, and the outcomes of medications. For instance, they frequently post the latest values of their blood glucose, diet, and exercises using "*#diabetes*" and "*#BGnow*" hashtags on Twitter, as show in Figure 3.1.

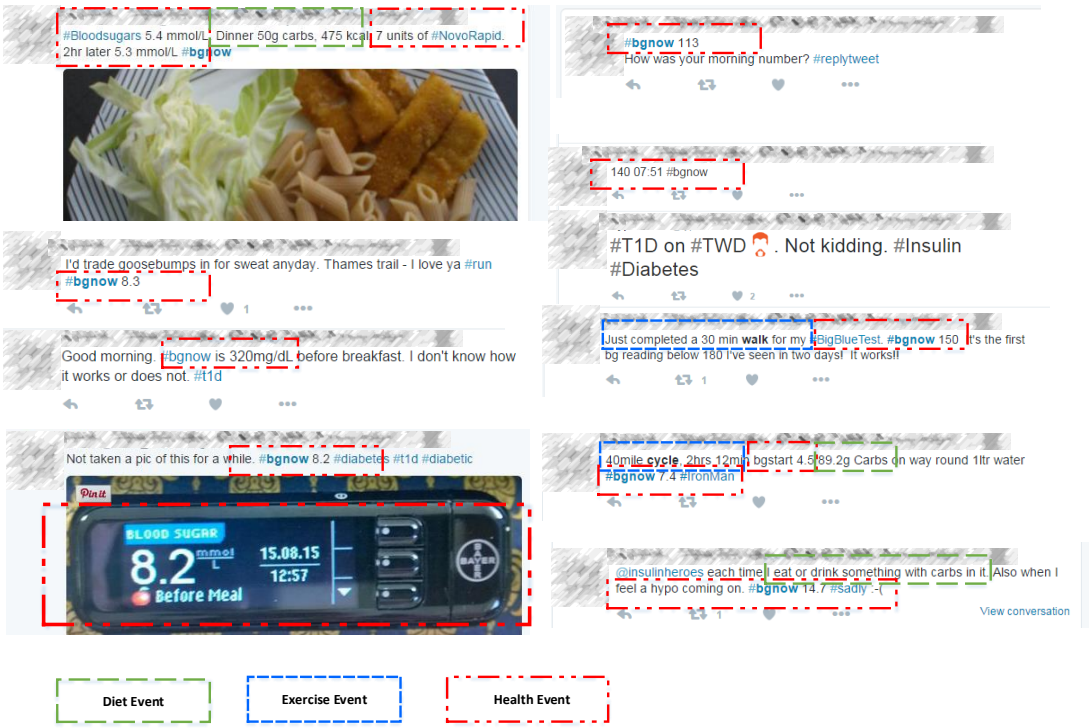


Figure 3.1: Examples of tweets which mention a personal wellness event.

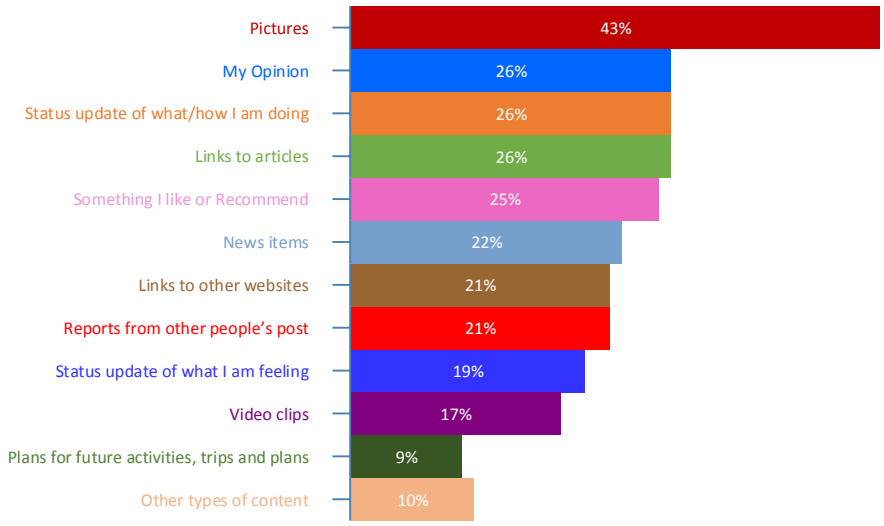


Figure 3.2: Most popular shared content on social media.

On the other hand, a survey conducted by Ipsos research center[1] in 2013 reported that updates pertaining to daily activities is ranked third in social media contents, as shown in Figure 3.2. In fact, social media platforms have become the most popular ways for users to share what happening around them. Hence, the abundance and growing usage of social media has made a large repository about users' daily happening and activities providing a stethoscope for inferring individual's lifestyle and wellness. This provides new opportunities to understand individuals' wellness that can be used to assist them in managing their health in a scope that previously was impossible.

Extraction of wellness information and events from users' published social contents is the key initial step towards understanding, modeling and predicting the wellness of users. This helps us filter out irrelevant information and content in social networks, and harvest relevant information for further analysis. As a first step towards accomplishing this end, in this chapter, we propose to extract tweets pertaining to the wellness of a user and categorize them into a wellness taxonomy which includes different categories such as diet event, activity event and health.

3.1 Motivation and Challenges

Extraction of personal wellness events (PWEs) will provide significant insights about individual's wellness and community lifestyle behaviours. At the individual level, it can summarize the past wellness events of individuals, which signifi-

[1] http://www.socialmediatoday.com/content/what-most-popular-content-shared-social-media

cantly facilitates lifestyle management through coarse and fine-grained browsing. Further, PWE summary can be useful for downstream applications such as user health profiling, personalized lifestyle assistant, and targeted online advertising. Take diet as an example; if one diabetic person consumes a lot of carbohydrates, the system can offer diet suggestion. At the community level, accumulating the wellness information of a large population of individuals makes it feasible to analyze and understand the lifestyle patterns and wellness of social groups in a scale that was impossible with traditional methods in terms of both time and cost.

Despite its value and significance, extracting PWEs from social media services has not been fully investigated due to the following challenges. First, the language used in social media is highly varied, informal, and full of slang words. Second, PWEs are relatively rare in social media posts as users tend to post their personal significant events together with lots of trivialities and other public events (Li et al., 2014). As a result, wellness events are buried among other contents produced by the users and their social connections. Identifying wellness events from a huge volume of other non-wellness events poses a big challenge. As a result, even a large annotated dataset might contain just a few examples of PWE categories. Third, the structure of wellness events exhibits a hierarchical taxonomy as shown in Table 3.1. Indeed, events under the same category are closely related. For instance, clinical tests are much more related to treatment, than running. These events may share some features such as entities, attributes and relations, which makes the problem arduous. How to mathematically model such relations and

integrate them into a learning framework remains a challenge.

3.2 Overview

In health sciences, it has been intensively studied and well-established that physical activities, diet planning and taking prescribed medications are the key therapeutic treatments of many diseases (Pastors et al., 2002, Hu, 2011). Further, unhealthy lifestyle behaviours such as unhealthy dietary habits, sedentary lifestyle, and the harmful consumption of alcohol are mainly related to the risk factors of noncommunicable diseases (NCDs) ranked as the leading cause of disability-adjusted life years (DALYs) (Lim et al., 2013, Association and others, 2014). Therefore, the primary aim of the General Assembly of the United Nations on NCDs in 2011 was to reduce the level of exposure of individuals and population to NCDs' risk factors and strengthen the capacity of individuals to follow lifestyle patterns that foster good health[2].

Motivated by this discussion, as a first step towards understanding and analyzing users' wellness from social media, we propose a supervised model to extract PWEs from social media posts of a given user and categorize them into a wellness taxonomy as shown in Table 3.1. In particular, we propose an optimization learning framework that utilizes the content information of microblogging messages as well as the relations among event categories. We seamlessly incorporate these two types of information into a sparse learning framework to tackle problems arising

[2]http://www.un.org/en/ga/ncdmeeting2011/

from noisy texts in microblogs.

The advantageous of our proposed method are multifold:

- We introduce an approach for harvesting wellness information available on social media for studying public health. In particular, we propose an effective framework for collecting users' wellness event from social media, which scales well. Although experiments were performed on diabetic users who use Twitter microblogging platform, it is easily extendable to other diseases. As far as we know, this is the first study on personal wellness event extraction from social media posts of individuals.

- We present a novel supervised model for wellness event extraction that takes task relatedness into account to learn task-specific and task-shared features.

- We construct a large-scale diabetes dataset by automatically extracting lifestyle and wellness related short messages and manually constructing the ground-truth labels.

3.3 Problem Statement

The problem we study in this paper is different from traditional event detection since the latter normally focuses on detecting and constructing an evolutionary timeline of public events (Becker, Naaman, and Gravano, 2011, Meladianos et al., 2015). Moreover, they assume that events are independent and hence only consider content information to identify event categories. In this section, we first

Table 3.1: Taxonomy of wellness events with exemplar tweets.

Event	Sub Event	Example tweet mentioning an event
Diet	Meals	Dinner just salad
	Alcoholic Beverages	Too much drink in party
	Non-alcoholic Beverages	Talking about hot chocolates, I might just go and make myself one :D
	Snacks	found Taylor's pretzels in my backpack and I'm so happy wow
	Fruit	almost eat all the strawberries
	Others	Eat 20g carbs and go fo running
Exercise	Walking	20 mins walk around office..
	Running	after 1 hour run #bgnow 130
	Biking	I just finished 1 hour biking
	Swimming	BGnow 95, thanks swimming pool
	Others	Shopping and having a little dinner URL
Health	Examinations	#BGnow 100
	Symptoms	Feel too much Fatigue
	Treatment	ate great oatmeal, toast, and eggs. Had 1 unit

present the notations and then formally define the problem of PWE detection from individuals' social media accounts.

Suppose that there are M wellness events and let $\mathcal{C} = \{c_1, c_2, ..., c_M\}$ be the set of class labels. Given a corpus $\mathcal{P} = \{p_1, p_2, ..., p_N\}$ composed of N different training samples. Each training sample $p_i = (\mathbf{x}_i, \mathbf{y}_i)$ consists of a message content vector denoted by $\mathbf{x}_i \in \mathbb{R}^J$ and the corresponding event label vector denoted by $\mathbf{y}_i \in \mathbb{R}^M$. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]^T \in \mathbb{R}^{N \times J}$ be the matrix representing training data and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]^T \in \mathbb{R}^{N \times M}$ be the matrix of labels. Our learning task is to find a mapping function from feature space $\mathbf{X}$ to label $\mathbf{Y}$.

With the notation above, we formally define the personal wellness event detection problem as: *Given a sequence of microblog messages $\mathcal{P}$ with their content $\mathbf{X}$, and the corresponding event labels $\mathbf{Y}$, we aim to learn a model $\mathbf{W}$ to automatically assign events' labels for unseen messages.*

3.4 Wellness Event Categorization

In essence, two characteristics of personal wellness event detection are: 1) training data is sparse; and 2) event categories are deeply inter-related. Their associated challenges are: a) which events are related in problem domain; and b) how to incorporate event relations into the learning framework to infer a more effective learning model. In this section, we first explain how to formulate the problem of PWE detection as a multi-task learning (MTL) framework which utilizes the content information of microblogging texts as well as captures the relations between the event categories into an integrated learning framework. We samelessly integrate these two types of information into a state-of-the-art framework and turn the integrated framework into an optimization problem. We then demonstrate how to find the solution of the problem with an efficient framework.

3.4.1 Modeling Content Information

Traditionally, supervised learning is widely used to infer topics of text documents. A straight forward way for event detection is to learn a supervised model based on labeled data, and apply the model to detect the topics of each microblogging post. However, compared with textual documents in traditional media, a distinct feature of texts in microblogging platforms is that they are noisy and short (Chen et al., 2013, Hu, Tang, and Liu, 2014), which give rise to two issues. First, text representation models, like "Bag of Words" (BoW) and n-grams, lead to a high-

dimension feature space due to the variety of words. Second, the posts are too short and noisy making the representation very sparse. To mitigate these problems, we propose a sparse model to perform classification of feature space.

Assume that we have M wellness events, and view each event as one task. Formally, we have M tasks $\{T_1, T_2, ..., T_M\}$ in the given training set $\mathcal{P}$. The prediction for each task t is given by $f_t(\mathbf{x}; \mathbf{w}_t) = \mathbf{x}^T \mathbf{w}_t$, where $\mathbf{w}_t$ is the coefficient for the task t. The weight matrix of all M tasks can be denoted as $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M] \in \mathbb{R}^{J \times M}$. Matrix $\mathbf{W}$ can be inferred from the training data by solving the following optimization problem:

$$\arg\min_{\mathbf{W}} \mathcal{L}(\mathbf{X}, \mathbf{W}, \mathbf{Y}) + \Phi(\mathbf{W}), \tag{3.1}$$

where $\mathcal{L}(.)$ is the loss function, and $\Phi(\mathbf{W})$ is a regularizer which controls the complexity of the model to prevent overfitting and selects discriminant features. This formulation is a sparse supervised method, where the data instances are independent and identically distributed (*i.i.d*), and the tasks are independent. There are several choices for the loss function, i.e., $\mathcal{L}(\mathbf{X}, \mathbf{W}, \mathbf{Y})$, in machine learning depending on the focused task. Two common choices are square loss and logistic loss (Rosasco et al., 2004). Logistic loss function can better handle the multi-label data as frequently reported in machine learning studies (Kumar and Daume III, 2012).

Upon the above discussion, the loss function $\mathcal{L}(\mathbf{X}, \mathbf{W}, \mathbf{Y})$ is defined as logistic

loss in this work,

$$\sum_{t=1}^{M}\sum_{i=1}^{N} log(1 + exp(-y_i^t f_t(\mathbf{x_i}, \mathbf{w_t}))), \tag{3.2}$$

where $y_i^t \in \{-1, 1\}$ is the true label indicating the relevance of i-th sample to the t-th task. Note that each sample can fall into multiple categories. For instance, "banana bread in the oven, mmmmm! lets just enjoy this #bgnow 70!" is related to meals and health examination categories at the same time. In this example, the user reported his blood glucose value, i.e., 70 , and his decision to eat some banana bread.

To select discriminant features and control the complexity of our model, we define $\Phi(\mathbf{W})$ as follows,

$$\Phi(\mathbf{W}) = \alpha\|\mathbf{W}\|_F^2 + \beta\|\mathbf{W}\|_1, \tag{3.3}$$

where, α and β are positive regularizer parameters. In the defined regularizer $\Phi(\mathbf{W})$, the first term, i.e. Frobenius-norm, controls the generalization performance of the model and the second term, i.e. ℓ_1-norm, leads to a sparse representation for the texts, which performs feature selection to reduce the effects of noisy features. Thus $\Phi(\mathbf{W})$ performs a kind of continuous feature selection as well as controls the complexity of the model (Ruvolo and Eaton, 2014, Song et al., 2015a).

3.4.2 Modeling Events Relations

Recall that PWE detection has two characteristics: 1) some events are more related to each other while differ from others, and similar events might share some features. For example, "walking" shares some features with "running" since the context of two events are similar. However, it greatly differs from "meals"; 2) the dimension of feature space is usually very high. In fact, some features are not discriminative enough for wellness event detection. This motivates us to propose a graph-guided multi-task learning model, which is capable of capturing the relatedness among tasks to learn task-shared features as well as the task-specific features. The hope is that common information relevant to prediction can be shared among tasks and joint learning of tasks' models leads to a better generalization performance as compared to learning each task independently. A major challenge hence is how to control the sharing of information among tasks so that it leads to close models for related tasks while unrelated tasks do not end up influencing each other. Therefore, the key assumption for our model is that tasks are assumed to be related to each other with different weights and the parameters of two related tasks are close to each other in ℓ_2 norm sense.

Based on the above discussion, to incorporate task relations into event detection, we assume that the task relationships can be represented using a graph structure G, where each node represents one task and each edge connects two related tasks. The weight of each edge $r(t_i, t_j)$ reflects the relation strength be-

tween two tasks i and j. Given a graph G, we can formulate the task relations as minimizing the following objective function $\Omega(\mathbf{W})$,

$$\begin{aligned}\Omega(\mathbf{W}) &= \lambda \sum_{t_i,t_j \in \mathcal{E}} r(t_i,t_j)\|\mathbf{W}_{*i} - \mathbf{W}_{*j}\|_2^2 \\ &= \lambda \operatorname{tr}(\mathbf{W}(\mathbf{V}-\mathbf{R})\mathbf{W}^T) = \lambda \operatorname{tr}(\mathbf{W}\boldsymbol{\Delta}\mathbf{W}^T),\end{aligned} \tag{3.4}$$

where $\mathcal{E}$ contains all the edges of graph G, and $\boldsymbol{\Delta} = \mathbf{V} - \mathbf{R}$ is the graph Laplacian matrix (Nie et al., 2014a, Akbari, Nie, and Chua, 2015), where $\mathbf{R} \in \mathbb{R}^{M \times M}$ is the task relatedness matrix. $\mathbf{R}_{ij} = r(t_i, t_j)$ indicates the relation strength between task i, and j and $\mathbf{R}_{ij} = 0$, otherwise. $\mathbf{V} = diag(\mathbf{V}_{jj})$ is a diagonal matrix with $\mathbf{V}_{jj} = \sum_{i=1}^{M} r(t_i, t_j)$. The regularizer parameter λ controls the impact of relations amongst tasks in the learning process.

To construct the graph, we utilize a fully automated approach based on the model learnt from the relaxed multi-task problem. Following the idea discussed in (Kim and Xing, 2009), we first train a MTL model with Lasso regularizer to compute the model for each tasks t_i and then compute the pairwise correlation between distinct tasks. We simply create an edge between each pair of tasks which have correlation above a defined threshold ρ. We set the threshold to $\rho = 0.7$ since it leads to the best performance in our experiments.

The optimization framework, which integrates content information and event relation information into the learning process, is defined by the integration of Eq.

(3.1), through Eq. (3.4) as the following objective function, $J(\mathbf{W})$,

$$\arg\min_{\mathbf{W}} J(\mathbf{W}) = \mathcal{L}(\mathbf{X}, \mathbf{W}, \mathbf{Y}) + \Phi(\mathbf{W}) + \Omega(\mathbf{W}), \tag{3.5}$$

where the first and second terms are to consider content information and perform regularization to avoid overfitting, respectively. The third term, i.e. $\Omega(.)$, captures tasks relatedness to learn task-shared features.

3.4.3 Optimization

The objective function $J(\mathbf{W})$ (i.e., Eq. (3.5)) is non-smooth since it is the composition of a smooth term and a non-smooth term, i.e. ℓ_1 penalty, and gradient descent method is not available to solve the formulation. In this section, we introduce an efficient algorithm to solve the optimization problem.

Inspired by (Nesterov, 2004, Chen et al., 2009), we propose to solve the non-smooth optimization problem in Eq. (3.5) by optimizing its equivalent smooth convex reformulation. We hence derive an smooth reformulation of Eq. (3.5) by moving the non-smooth part, i.e. ℓ_1 norm, to the constraint.

Lemma 3.1. $\|\mathbf{W}\|_1$ *is a valid norm.*

Proof. It is easy to verify that $\|\mathbf{W}\|_1$ satisfies the three conditions of a valid norm, including the triangle inequality $\|A\|_1 + \|B\|_1 = \|A + B\|_1$, which completes the proof. □

Theorem 3.1. *Let* $\mathcal{L}(\mathbf{X}, \mathbf{W}, \mathbf{Y})$ *be a smooth convex loss function. Then Eq. (3.5)*

can be reformulated as the following ℓ_1-ball constrained smooth convex optimization problem:

$$\arg\min_{\mathbf{W}\in\mathbf{Z}} f(\mathbf{W}) = \mathcal{L}(\mathbf{X}, \mathbf{W}, \mathbf{Y}) + \lambda \operatorname{tr}(\mathbf{W}\boldsymbol{\Delta}\mathbf{W}^T) + \alpha\|\mathbf{W}\|_F^2, \tag{3.6}$$

where,

$$\mathbf{Z} = \{\mathbf{W}|\|\mathbf{W}\|_1 \leq z\}, \tag{3.7}$$

$z \geq 0$ is the radius of the ℓ_1-ball and there was a one-to-one correspondence between β and z.

Proof. We first prove that Eq. (3.6) is a constrained smooth convex optimization problem. From the Lemma 3.1, we know that $\|\mathbf{W}\|_1$ is a valid norm. In (Nesterov, 2004) Nesterov proved that any norm is a closed convex function so we can conclude that $\|\mathbf{W}\|_1$ is a closed convex function. We hence conclude that $\mathbf{Z} = \{\mathbf{W}|\|\mathbf{W}\|_1 \leq z\}$ is a closed and convex set (Theorem 3.1.3, (Nesterov, 2004)). It is easy to verify that the objective function $f(\mathbf{W})$ is convex and differentiable since it is the composition of convex functions.

As we can see, our problem defines a convex and differentiable function $f(\mathbf{W})$ in a closed and convex set $\mathbf{Z}$. Thus this problem is a constrained smooth convex optimization problem and the equivalence of Eq. (3.5) and Eq. (3.6) follows from the Lagrngian duality which completes the proof. □

We now find the solution for Eq. (3.6), which is equivalent to our optimization problem in Eq. (3.5). To solve the problem, we first consider the optimization problem without the constraint on $\mathbf{Z}$ which is defined as:

$$\arg\min_{\mathbf{W}} f(\mathbf{W}). \tag{3.8}$$

The solution to this problem can be computed from the gradient descent method which iteratively updates $\mathbf{W}_{i+1}$ in each step as follows,

$$\mathbf{W}_{i+1} = \mathbf{W}_i - \frac{1}{\gamma_i}\nabla f(\mathbf{W}_i), \tag{3.9}$$

where γ_i is the step size and it is determined by line search according to Armijo-Goldstein rule (Nesterov, 2004). The smooth part of the optimization problem can be reformulated equivalently as a proximal regularization of the linearized function $f(\mathbf{W})$ at $\mathbf{W}_i$ as,

$$\mathbf{W}_{i+1} = \arg\min_{\mathbf{W}} \mathcal{M}_{\gamma_i,\mathbf{S}_i}(\mathbf{W}), \tag{3.10}$$

where,

$$\mathcal{M}_{\gamma,\mathbf{S}_i}(\mathbf{W}) = f(\mathbf{S}_i) + \langle \nabla f(\mathbf{S}_i), \mathbf{W} - \mathbf{S}_i \rangle + \frac{\gamma_i}{2}\|\mathbf{W} - \mathbf{S}_i\|_F^2, \tag{3.11}$$

where $\mathbf{S}_i$ is computed based on the past solutions by $\mathbf{S}_i = \mathbf{W}_i + \tau_i(\mathbf{W}_i - \mathbf{W}_{i-1})$.

Eq. (5.17) can be rewritten as,

$$\arg\min_{\mathbf{W}} \mathcal{M}_{\gamma_i,\mathbf{S}_i}(\mathbf{W}) = \tag{3.12}$$
$$\arg\min_{\mathbf{W}} (\frac{1}{2}\|\mathbf{W} - (\mathbf{S}_i - \frac{1}{\gamma_i}\nabla f(\mathbf{S}_i))\|_F^2)$$

By ignoring terms that are independent of $\mathbf{W}$ the objective function boils down to:

$$\mathbf{W}_{i+1} = \arg\min_{\mathbf{W}} \|\mathbf{W} - \mathbf{U}_i\|_F^2, \tag{3.13}$$

where $\mathbf{U}_i = \mathbf{S}_i - \frac{1}{\gamma_i}\nabla f(\mathbf{S}_i)$ and indeed the solution of $\mathbf{W}$ is the Euclidian projection of $\mathbf{U}_i$ on $\mathbf{Z}$. Upon this discussion, we can have an efficient and optimal solution to the convex optimization problem. Similar to the proof in (Liu, Ji, and Ye, 2009a), it is easy to show that the convergence rate of the proposed algorithm is $O(\frac{1}{\epsilon})$ for achieving an accuracy of ϵ. The overall optimization process can be described in Algorithm 3.1.

In the algorithm, we utilize Nesterov's method (Nesterov, 2004) to solve the optimization problem in Eq. (3.5). We use the line search algorithm for γ_i from line 5 to line 11 according to the Armijo-Goldstein. In line 12, t_i is set according to (Liu, Ji, and Ye, 2009a). Based on the algorithm, we can compute the solution for the convex optimization problem.

Algorithm 3.1: Optimization algorithm of Eq. (3.5)

Input: $\mathbf{W}_0$, $\gamma_0 \in \mathbb{R}$, and q = max iteration.
Output: $\mathbf{W}$.

Set $\mathbf{W}_1 = \mathbf{W}_0, t_{-1} = 0$, and $t_0 = 1$.
for $i = 1$ *to* q **do**
  Set $\tau_i = (t_{i-2} - 1)/t_{i-1}$.
  Set $\mathbf{S}_i = \mathbf{W}_i + \tau_i(\mathbf{W}_i - \mathbf{W}_{i-1})$.
  while *true* **do**
    Compute $\mathbf{W}^* = \arg\min_{\mathbf{W}} \mathcal{M}_{\gamma_i,\mathbf{S}_i}(\mathbf{W})$
    where $\mathcal{M}_{\gamma,\mathbf{S}_i}(\mathbf{W}) = f(\mathbf{S}) + \gamma\|\mathbf{W}\|_{1,1} + \langle\nabla f(\mathbf{S}), \mathbf{W} - \mathbf{S}\rangle + \frac{l_i}{2}\|\mathbf{W} - \mathbf{S}\|_F^2$
    if $f(\mathbf{W}^*) \leq \mathcal{M}_{\gamma_i,\mathbf{S}_i}(\mathbf{W})$ **then**
      break
    else
      Set $\gamma_i = \gamma_i \times 2$
  Set $\mathbf{W}_{i+1} = \mathbf{W}^*$ and $\gamma_{i+1} = \gamma_i$. Set $t_i = \frac{1+\sqrt{1+4t_{i-1}^2}}{2}$.
Set $\mathbf{W} = \mathbf{W}_{i+1}$.

3.5 Experiments

In this section, we present the experimental details to verify the effectiveness of the proposed framework. We conduct experiments to answer the following questions that help to validate the framework:

1. How does the proposed framework perform as compared to other state-of-the-art baselines?

2. How well the selected features discriminate PWEs?

3. How sensitive is our model to the involved parameters?

In the rest of the section, we first introduce the dataset and experimental settings. We then respectively explore the answers to the aforementioned experimental questions. We finally summarize the key findings from the experiments.

3.5.1 Dataset Description

Recall that one of the main problems of this research is the lack of training data. According to our statistics, the wellness-oriented tweets are only less than 5% of all the messages posted by the chronic disease sufferers, and this value could be much smaller for healthy users. Therefore, we utilize a bootstrapping method to harvest the tweets corresponding to wellness events. We then manually label this tweet pool to construct our ground truth.

Wellness event categories. As mentioned perviously, it has been intensively studied and well-understood that physical activities, diet planning and taking prescribed medications are key therapeutic treatments of many diseases (Pastors et al., 2002). Inspired by (Shelley, 2005, Teodoro and Naaman, 2013), we arrive at three high-level wellness categories, namely, diet, exercise & activities (exercise for brevity), and health as shown in Table 3.1. Under each high-level event category, we further organize specific sub-events which construct a taxonomy comprises 14 distinct wellness events. We also define a null class for non-wellness events indicating that a message is not directly related to any defined wellness event categories.

Assigning event labels. We observed that different wellness events place emphasis on different hashtags and words. For instance, we observed that "#dwalk" regularly appears in walking related posts. Inspired by (Mintz et al., 2009, Gupta and Manning, 2014), we adopted a bootstrapping approach to select a set of tweet

related to each wellness event. To do so, we first selected some representative seed words for each wellness events by verifying top frequent keywords of each category. We then gathered tweets explicitly involving these seed words. However, the collected tweets are weakly related to events and are full of noises. For instance, the tweet *"I love music,it has a voice for every walk of life,every emotion,every bit of love"*[3] even containing the word "walk", but it is not a relevant one. To filter irrelevant tweets, we defined patterns in local context of each seed word. We applied the bootstrapping approach of (Thelen and Riloff, 2002) to extend the set of keywords and patterns and collected more positive samples pertaining to wellness events. We stopped bootstrapping after 10 iterations since it often failed to find more positive candidates.

To construct the dataset, we first crawled a set of users who used #BGnow hashtag in their tweets. This hashtag is very popular among diabetic patients to post information about diabetes and their health states. In this way, we gathered $2,500$ different diabetes users. We removed accounts which had high daily traffic to avoid spammers. This filtering process resulted in $1,987$ diabetic users. We then crawled all historical tweets of these users using Twitter API, resulting in a set of about 3 million tweets. We applied the aforementioned bootstrapping procedure to find candidate tweets to construct dataset, which resulted in $11,217$ tweets. We manually labelled all the tweets based on the wellness events as shown in Table 3.1. For each given event, we regarded tweets labelled with its class as

[3]This is a real tweet from the dataset.

the positive training samples, and randomly selected negative samples from other events. Examples of the positive and negative tweets for the event "walking" are given below:

Positive 3 litres of water and 4 miles of walking I am feeling super refreshed...thank god!!

Negative Further evidence of the benefits of exercise for people with type 2 #diabetes URL #doc (Error: It is not an event but reports general health information).

Table 3.2 shows the statistics of our dataset. In total, our training set consists of approximately $3,000$ tweets corresponding to different wellness events. We also randomly selected about $3,000$ non-wellness tweets to be used as positive samples for the null class (non-wellness events). We intentionally selected more samples for null class due to the imbalance nature of events. We divided the dataset into two sets based on their posting times. In particular, tweets that were posted before May 2015 were utilized to train our model; while those posted from May to July 2015 were used for evaluation process[4]. We call this dataset as **BG dataset** throughout this dissertation as it is constructed based on the hashtag "#BGNow".

We engaged another annotator to manually examine about $3,000$ messages. The inter-agreement between annotator was 0.857 with the *Cohen* κ metric, which verifies a substantial agreement between annotators.

[4]Note that the numbers in Table 3.2 do not add up to $11,217$ since our dataset is a multi-label dataset meaning that some messages discuss about more than one PWE.

Table 3.2: Statistics of the BG Dataset.

	All samples	Positive samples
Posts on Diet	1979	710
Posts of Exercise	2771	1234
Posts on Health	8802	1300
Total Number of Posts	11,217	3,244

3.5.2 Feature Settings

Content and linguistic features are two major features which are used for text classification (Cherry, Guo, and Dai, 2015, Moens, Li, and Chua, 2014, Hu and Liu, 2012). We follow them and extract the following set of features to represent each tweet from both context and linguistic aspects:

- **NGrams**: We extracted unigrams and bigrams from Twitter messages since they are commonly used to represent user-generated contents (Hu and Liu, 2012).

- **NE**: As shown in (Li et al., 2014), the presence of named entities is a very useful indication of events in social media texts. We hence utilized named entities as another feature to represent tweets (Ritter et al., 2012).

- **Gazetteer**: Gazetteers are commonly used as a linguistic feature in domain specific applications (Carlson, Gaffney, and Vasile, 2009). Hence, we used a dictionary of popular food and drink names from (Abbar, Mejova, and Weber, 2015) to extract gazetteer feature for foods and drinks. We also utilized LIWC's time category which includes 68 time terms (Pennebaker, Francis, and Booth, 2001).

Table 3.3: Performance comparison among models.

Method	Precision	Recall	F-1 score
Alan12	62.70	48.10	54.44
SVM	83.05	79.65	81.31
Lasso	80.45	79.21	79.82
GL-MTL	84.37	80.72	82.50
TN-MTL	83.22	78.85	80.98
gMTL	**87.15**	**82.69**	**84.86**

- **Modality**: Twitter has evolved to become a general purpose platform for communication. Therefore, users often share general thoughts, wishes, and opinions in their account. For our purpose of understanding personal wellness events, we need to filter out these irrelevant information from those which really report an event. Modality speeches has been used to express the possibility or uncertainty of events (Li et al., 2014). Hence, we utilized modality verbs as an indicator of non-event information. We check whether the message includes some modality verbs such as “may”, “could”, “must” and etc.

3.5.3 On Performance Evaluation

We conducted experiments to compare the performance of our model with other state-of-the-art approaches:

- **Alan**12: Event extraction method of (Ritter et al., 2012) which learns a latent model to uncover appropriate event types based on available data.
- **SVM**: We trained a binary classifier for each event to infer the label of

tweets.

- **Lasso**: Logistic regression model with Lasso regularizer, i.e. ℓ_1 term (Tibshirani, 1996).
- **GL-MTL**: Group Lasso regularizer with $\ell_{1/2}$ norm penalty for joint feature selection (Nie et al., 2010), which only encodes group sparsity.
- **TN-MTL**: Trace Norm Regularized MTL (Obozinski, Taskar, and Jordan, 2010), which assumes that all tasks are related in a low dimensional subspace.
- **gMTL**: Our proposed wellness event detection model.

For each method mentioned above, the respective parameters were carefully tuned based on 5-fold cross validation on the training set and the parameters with the best performance were used to report the final comparison results. The overall performance is shown in Table 3.3 in terms of precision, recall, and F-1 score metrics.

From the table, we can observe that all MTL methods outperform **Alan**12, **SVM** and **Lasso** in terms of precision with a substantial improvement over **Alan**12. The main reason is that event discovery methods mostly focus on detecting general events or major personal events (Zhou, Chen, and He, 2015). These events are discussed bursty and highly connected to specific name entities such as organizations, persons, and locations. However, PWEs are merely focus of individuals' local circles and may not be significantly related to any specific name entities. This hinders the learning framework to find representative latent topics from

data. Among the multi-task approaches, **gMTL** achieves the best performance as compared to others. It verifies that there exists relationships among events and such relatedness can boost the learning performance. **GL-Lasso** achieves higher performance as compared to **Lasso** and **TN-MTL** since it tries to jointly learn features which resulted in better generalization. This verifies that sharing samples among distinct task alleviates the data scarcity problem as pointed out by previous studies (Ruvolo and Eaton, 2014, Xu et al., 2015). The proposed **gMTL** model outperforms other methods by 2%-6% since it encodes the task relatedness and group sparsity. By sharing samples between different tasks, i.e. event categories, **gMTL** simultaneously learns task-shared and task-specific features as well as mitigates the problem of data scarcity.

3.5.4 On Feature Comparison

We also conducted an experiment to evaluate the effects of different features for PWE detection, as shown in Table 3.4. To conduct the study, we considered **NGram** feature as a baseline feature since it has been shown in many studies to have good performance (Tang et al., 2014, Nie et al., 2014b). We then added each distinct feature from the feature set and reported the average performance over all event categories. We also performed significant test to validate the importance of different features. We used the asterisk mark (*) to indicate significant improvement over the baseline.

As Table 3.4 shows, **NGram** and **Gaz** are important features for PWE de-

Table 3.4: Average performance of PWE detection on different feature setting.

Method	Precision	F-1 score
NGrams (Baseline)	82.70	81.06
Baseline+Gaz *	84.31	82.31
Baseline+Gaz+NE	86.85	84.04
All *	87.15	84.86

tection. The reason might be that **NGram** represents the context information of messages and food gazetteer feature is a very effective indicator of events related to food and drink category which filter out many irrelevant samples. However, adding name entities, i.e. **NE**, improves the performance but not significantly. This shows that this feature may not be effective for wellness event detection, as we had expected, though it is widely used for public event detection. We also observed that **Modality** feature is useful for event detection. Indeed, we observed that it is able to filter out activities from wishes or general thoughts and information significantly.

3.5.5 On Parameter Sensitivity

An important parameter in our method is λ in Eq. (3.4) that determines the impact of relation amongst tasks in the learning process. A high value indicates the importance of these relations while a low value limits the effect of relations amongst tasks. Another important parameter is the number of selected features. Hence, we study how the performance of our model varies with λ and the number of selected features. Figure 3.3 shows the performance of our model with different parameter settings which achieves the peak of 84.86% when $\lambda = 0.01$ and 1400

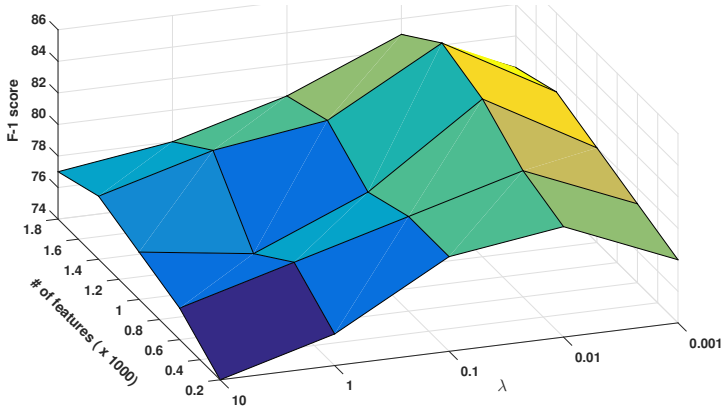


Figure 3.3: The impact of different parameter setting.

features was selected. The general pattern is that the performance is more sensitive to the number of selected features, and the best number of features is around 1400; furthermore, there is not a significant improvement above this point. It is worth noting that how to determine the number of features is still an open problem in data mining (Wang, Tang, and Liu, 2015).

3.6 Related Work

It is worth emphasizing that several practical systems and research efforts have been dedicated to event extraction from Twitter (Zhao et al., 2013a, Diao et al., 2012, Li and Cardie, 2014), however, existing approaches mostly focus on public events (e.g. disaster outbreak) or major personal events (e.g. marriage, job) which are often discussed significantly in social media platforms. The bursty nature of these events allows the detection algorithm to utilize a large amount of aggregated data to detect bursts or irregularities in topical contents in a short time span (Li and Cardie, 2014, Li et al., 2014). PWEs, in contrast, are not highly discussed

in social media since they might only be important for the user himself or his close friends, i.e. his local circles. For example, message like "35 degrees, 25mph winds, and rain...it's gonna be a fun ride home...3.5 miles uphill :)" intuitively shows an exercise event but it might not be discussed too much in user's social account. Further, existing event extraction approaches assume that distinct events are independent from each other. However, previous studies in text mining have proven that events are not independent and events under the same category are closely related (Gella, Cook, and Baldwin, 2014).

3.7 Summary

Personal wellness events, in contrast to public events in social platform, are rarely discussed and deeply related to each other. In this study, we proposed a learning framework that utilizes content information of microblogging texts as well as the relation between event categories to extract PWE from users social posts. In particular, we modeled the inter-relatedness among distinct events as a graph Laplacian which was employed as a regularization to a sparse learning framework. Thus the proposed model not only can learn task-shared and task-specific features but is also robust to noise in microblogging contents. Experimental evaluations on a real-world dataset from Twitter revealed that our proposed framework significantly outperforms the representative state-of-the-art methods.

CHAPTER 4

Characterization Study of Diabetes on Twitter

With the increasing popularity of social media platforms, health consumers increasingly utilize social platforms to fulfill their health demands through seeking and sharing health information and experiences as well as providing online social support for their peers (Attai et al., 2015, De Choudhury, Morris, and White, 2014, Dredze, 2012). In essence, social environments and virtual communities have been transformed to a confident environment permitting users to be connected with their peers who have experienced similar conditions, difficulties, and challenges, assisting them to cope with their situations (Davis, Anthony, and Pauls, 2015, Greene et al., 2011). The emerging of self-tracking gadgets and the enthusiasm of users in taking informed health decisions has also intensified this trend. This motivates users disclose their health information in social platforms (De Choudhury et al., 2013). Further, the ubiquity of social media encourages health consumers to not only discuss about their health conditions and share experiences but more importantly share their health related measurements, like blood

pressure and blood glucose, which provides an invaluable resource to study and analysis individuals' and communities' wellness and behaviour (See examples in Chapter 1 and Chapter 3).

In previous chapter we demonstrated a learning framework for extracting wellness information of users from online social networks. In this chapter we aim to analyze the rich collected data to gain better insights about a specific wellness problem. To be more specific, we aim to study the capability of social media as a sensor for reflecting wellness statuses and attributes of social users. To do so, we take diabetes as an example in the wellness domain and investigate users' online behaviors to distinguish the success of users in managing their health condition. Through this study, we attempt to distinguish users who are successful in controlling their blood glucose from those who fail to do so.

4.1 Motivation and Challenges

Diabetes is the sixth leading cause of death in the US and it is estimated to be the seventh cause of the worldwide by 2030 [1]. It causes serious complications and can lead to poor quality of life (Association and others, 2014). People with diabetes are more suspectable to other illnesses. However, they can reduce the occurrence of these potential complications through diabetes self-management. Training and education of self-management of diabetes prevent unnecessary health care utilization and hospitalization and improve patients wellness (Haas et al., 2013).

[1] `http://www.cdc.gov/diabetes/data/statistics/2014statisticsreport.html`

However, diabetes self-management is not a trivial task for chronic patients since it is highly linked to several individual and cultural factors such as demographic attributes, psychological mood, patient knowledge and lifestyle. The prominent role of self-management in diabetes has resulted into the emergence of several diabetes management programs and schemes (Haas et al., 2013). Despite the success and efficacy of these programs, many factors affecting on the success of patients in managing their condition are still unclear or partially opaque due to the following reasons (Garrett and Bluml, 2005, Lorig et al., 2010). First, most of these studies have been performed in a controlled clinical setting and only investigate some of health factors of interest like diet and activities. Even though these factors are important the effects can vary from patient to patient making the reported results almost restricted to the local community studied. Second, the adherence of the user is a major factor in these programs. However, users adherence and ambition vary from time to time making the generalization of results optimistic. Third, traditional studies in health sciences are mostly based on an observational study or based on survey data from patients which are limited in the number of subjects and the time period of the study, making it hard to derive a comprehensive conclusion. Therefore, the factors affect on success of users in managing their diabetes are still not fully studied. In particular, the role of personal characteristics like user's behavior and mood is still unclear.

4.2 Aim and Significance of the Study

In this chapter, we investigate the success of users to self-manage their diabetes, the role of users' behaviors on self-management, and how exactly user's behaviors correlate to self-management and health of the user. In particular, we investigate the factors which differentiate two cohorts of users: Adapted and Non-Adapted diabetes patients, where they are divided by their success in controlling their blood glucose values on the healthy target range. In contrast to traditional studies, we utilize users' self-reported values of blood glucose on Twitter microbloging service. We study users from three aspects, namely, linguistic attributes, published textual content, and shared multimedia contents, i.e., images on Instagram, to find their behavioural distinction. More specifically, we address the following research questions:

RQ1: What are the characteristics which differentiate adapted and non-adapted users in terms of linguistic, cognitive and affective styles and attributes?

RQ2: How do adapted users differ from non-adapted users in terms of their posts content published in social platforms?

RQ3: How different are the multimedia interests of adapted and non-adapted users on social platforms, i.e. what entities and concepts do they often share in pictures on social networks?

Studying behavioural characteristics of successful patients provides significant insight about individual's wellness. First, understanding behaviour factors correlated with success in managing diabetes will help us to design proper intervention program to assist users coping with their condition and improve their wellness. Second, by understanding success factors in self-management, we can assist new diagnosed users as well as users struggling with their condition to adapt their lifestyle with the disease condition. Third, several years of research in user behaviour and wellness have clearly revealed that social networks and health are closely related through the social reinforcements come from observing and modeling others' behaviors (Ruths, Pfeffer, and others, 2014). By linking users in different categories, i.e. suggesting successful users to strugglers, we can assist them to learn positive behaviours from new friend on social communities (Maibach and Cotton, 1995,Korda and Itani, 2013,Hawn, 2009). This study is also beneficial for understanding collaborative behaviour of communities. By aggregating behaviour of individuals, we can shed light on public health, trying to understand the wellness patterns and trends of user groups in large scale. Overall, the answers to these questions will enhance our understanding of users' online interactions about diabetes, in particular on the Twitter microblogging platform and Instagram multimedia sharing service, and how their behaviours are correlated with their health condition, assisting us in various applications like group wellness analysis (Tang, Wang, and Liu, 2011), public health (Hawn, 2009,Luxton, June, and Fairall, 2012), and health policy making (Murray and Lopez, 1996), to name a few.

The contribution of this study are multifold:

- We conduct a behavioural analysis of social media users who can successfully manage their diabetes and the factors which cause a fail.
- We conduct an open vocabulary analysis that captures language use of diabetic patients toward identifying behavioural patterns result in self-managing diabetes.
- We investigate media contents shared by diabetes patients to reveal their preferences in sharing various visual concepts.

Our results open a new research direction to study wellness in large scale utilizing social media data. Further, the study demonstrates the potential of social media to design proper intervention and treatment programs for diabetic patients.

4.3 Data Collection and Ground-Truth Generation

4.3.1 Dataset

We utilized the **BG** dataset introduced in Chapter 3 (See Section 3.5.1) for conducting this study. As previously mentioned, the dataset was constructed based on a set of users who actively share their wellness information on Twitter and

Instagram. These users, besides posting about their diet, activities, and wellness, disclose their health information in terms of medical events and measurements like the onset of hypoglycaemia/heyperglocemia and their blood glucose values. A main characteristic of this cohort of users is that they report the exact values of their blood glucose in tweets using '#BgNow' hashtag. This hashtag is popular among diabetic patients to report their blood glucose values and other diabetic related information online. In essence, '#BgNow' plays the role of an online support group for diabetic patients through which they share, explore, and validate their health states and knowledge. The hashtag is special interest for us since it enables us to study the correlation between diabetes patients' behaviours and their health indicators, e.g. blood glucose value. Indeed, #BgNow hashtag acts as a social sensor through which we can measure the blood glucose values of individuals on social networks.

Analyzing this collection of tweets provides invaluable insight about diabetic patients and their health information disclose on social media as well as the correlation between their behaviour online and their health conditions. To link the twitter account of users with their Instagram, we obtained the cross-platform links in which they publish a post from their Instagram account on Twitter. By storing all such links, we can find their Instagram account and avoid the problem of user identification across different networks(Vosecky, Hong, and Shen, 2009). It is worth noting that, by utilizing this approach, we may fail to find the corresponding Instagram account for those users who are not post cross network, resulting

Table 4.1: Statistics of the **BG** Dataset

	Twitter	Instagram
# of Users	1,174	113
# of AC users	436	36
# of NC users	738	77
# of Posts	1,060,105	2,623
# of Bgnow posts	20,079	–
Avg. posts per users	903	17

into a smaller dataset for Instagram. Table 4.1 shows statistics of our datasets for both Twitter and Instagram.

4.3.2 Extraction of Blood Glucose Values

The aim of our research is to investigate the behavioural distinction between two different cohorts of patients with respect to their health attributes, in our study blood glucose values. To do so, we need to extract the reported measurements of blood glucose values in their tweets. Several approaches have been proposed in information extraction to detect the right piece of information from a text corpus, like pattern-based methods, supervised classification, and Conditional Random Field (CRF) (Chang et al., 2006). Here, we utilized a simple but effective rule based approach. Intuitively, we defined a set of regular expression to extract the measurement values of blood glucose for individuals as shown in Table 4.2. We followed a bootstrapping approach similar to (Thelen and Riloff, 2002) to ensure the coverage and diversity of the used patterns, where all extracted patterns are manually verified to ensure accuracy. Given a user post, we apply these set of

Table 4.2: Representative examples of regular expressions for extracting blood glucose values from users' posts.

195.0 BG @ 08:20AM after bike ride 90 minutes	NUMBER (mg—mmol) (BG—Bgnow)	BG : 195
Going on a 3 mile run, #BGnow 120 #bigbluetest	* (BG—BGnow) NUMBER	BG : 120

rules to find whether a given tweet contains any reported values of blood glucose[2].

4.3.3 Ground-truth Generation

Medical studies suggest that diabetic users maintain their blood glucose between 70 and 130, which is considered as controlled blood glucose. A measurement in this range demonstrates that the patient could successfully manage his blood glucose while he was unsuccessful otherwise. The out of range values normally need to be corrected by lifestyle changes or treatments. Motivated by this principle, we partition the users into two distinct cohorts based on their reported blood glucose values: **Adapted Cohort** (**AC**), and **Non-adapted Cohort** (**NC**). The definitions of these two cohorts are as follows,

- **Adapted Cohort (AC)**: An adapted user is able to control and maintain values of his blood glucose in the suggested range most of the times. For such a user, the probability of observing a blood glucose value in the safe range is more than t, where t is a predefined threshold.

- **Non-adapted Cohort (NC)**: A user is non-adapted if he fail to have a

[2]In some cases, several numerical values might be found as a candidate for blood glucose value. We used the value which is closer to #Bgnow hashtag as the reference value.

controlled blood glucose, i.e., his blood glucose measurements are usually out of range.

We intuitively set the threshold $t = 0.5$ to divide the users in our dataset into two groups of different blood glucose patterns. Mathematically, we utilized the following decision function to construct ground truth labels,

$$d(u_i) = \begin{cases} +1 & \text{if} \quad Pr(u_i \in AC) \geq t \\ \\ -1 & \text{otherwise} \end{cases}, \tag{4.1}$$

where $Pr(u_i \in AC)$, and $Pr(u_i \in NC)$ are the probability that the measurements for the user u_i are in the controlled range and out of the controlled range, respectively. Here The probability of having a measurement in the prescribed range is computed based on the history of his blood glucose values, which can be computed by the following formulation,

$$Pr(u_i \in AC) = \frac{\text{\# of reported values in prescribed range for user} \quad i}{\text{total \# of reported values for user} \quad i} \tag{4.2}$$

This grouping is coarse, but it is motivated by health studies (Franciosi et al., 2001, Malanda, Bot, and Nijpels, 2013) stating that users who can manage their blood glucose will have a better long term health and fewer diabetes complications. In the future, we aim to define more detailed groups, e.g., users who have on target, below target, above target measurements and also different trends like stable, and fluctuating trends. By using Eq.(4.1), we can intuitively divide our

diabetic patients into two groups of users based on the self-reported values of blood glucose extracted in Section 4.3.2, which clearly show how they have managed their diabetes. In the remainder of the chapter, we study how these two communities' online behaviour differ in terms of linguistic, textual, and visual content published on their social network posts.

4.4 Analysis Method

In this section, we explain different analytical experiments we applied for understanding the behavioural distinction between AC and NC users.

4.4.1 LIWC Analysis

To identify and understand behaviour distinction of adapted and non-adapted groups, we leverage a variety of indicators including linguistic and non-linguistic indicators. The motivation behinds the investigation is that several psychological studies demonstrate behavioural expression of individuals and their responses expose their life context, crises, and vulnerabilities (Pennebaker, Francis, and Booth, 2001). Our analysis, in this section, is largely based on LIWC, which has been widely used in literature to study individuals behaviours in depression (De Choudhury, Counts, and Horvitz, 2013c), addiction recovery (MacLean et al., 2015), anorexia (De Choudhury, 2015), to name a few. We hence examine three categories of attributes named (1) affective attributes, (2) cognitive attributes and (3) linguistic and stylistic attributes.

Affective attributes. Affective measures have attracted a lot of research in text and opinion mining to detect the objectivity of user towards products, organization, and services (Liu, 2012). Recently, affective measures have been largely utilized to measure emotional disclosure of users in social media (Kumar et al., 2015). Motivated from prior literature, we measure positive affect (PA) and negative affect (NA) based on LIWC categories. We also compute four other emotional expression indicators: anger, anxiety, sadness, and swear.

Cognitive attributes. Several studies in psychology have demonstrated that cognitive process is largely associated with health improvement. For example, greater usage of cognitive words is related to less anxiety after treatments (Alvarez-Conrad, Zoellner, and Foa, 2001). Cognitive words are also utilized for explanatory purposes and demonstrate the demands of individuals for understanding the situation. We therefore evaluate the cognitive process of individuals based on cognition and perception word categories of LIWC.

Linguistic attributes. We consider five measures of linguistic style: (a) Lexical Density: consisting of words that are verbs, adjectives (identified using NLTK's POS tagger), and adverbs. (b) Temporal References: consisting of past, present, and future tenses. (c) Social/Personal Concerns: words belonging to family, friends, social, work, health, and death. (d) Interpersonal Awareness and Focus: words that are 1st person singular, 1st person plural, 2nd person, and 3rd person pronouns. (e) We also evaluated words associated with quantities such as numeric values as diabetes patients frequently need to consider amount of and quantities

of their foods, medications and activities to manage their health condition.

4.4.2 Topical Content Analysis

We studied the textual contents shared by diabetic users from two aspects: words and phrases (N-grams), and topics discussed.

N-Gram Analytic. In addition to linguistic analysis, we also investigate the usage of various n-grams in the contents shared by adapted and non-adapted people. Specifically, we investigate to discover the difference in usage of uni-, bi-, and tri-grams between two groups[3]. However, comparison between two set of n-grams is a challenging task mostly demonstrated by word-cloud. Inspired by recent research efforts in computational social sciences (Kumar et al., 2015), we compute the log-likelihood of the ratio between usage pattern of each n-gram between adapted and non-adapted groups. Mathematically, it can be computed as follows,

$$LLR = 2 \times [\ln \Big(Pr(u_i \in AC)\Big) - \ln \Big(Pr(u_i \in NC)\Big)] \tag{4.3}$$

Indeed, LLR demonstrates a clear measure to compare the differences between usage of an n-gram between two groups. As it computes the log likelihood ratio, when a n-gram is equally used in two groups then its LLR will be near zero. Meanwhile, it would be greater than zero if it is more frequent in first group as

[3]In this chapter, we use the general term 'n-gram' to refer to uni-, bi-, and tri-grams in the text.

compare to the second group, whereas it would be less than zero if the pattern is reverse [4].

Topic Analytic. Although content analysis based on n-grams provides an intuitive way for understanding the published contents by individuals, it processes the text based on low level features, i.e., words, and may fail to capture high level semantics inside the text. We hence apply topic models to discover the semantic topics inside the posts published by different user groups. Topic models have been commonly used to analyze health data (Paul and Dredze, 2014). Following the prior literature, we obtain topics by applying latent Dirichlet Allocation (LDA) over the entire set of posts shared by all users. To train the topic model, we used the default hyper-parameter settings and set the number of topics to 50, which we observed to work well in our experiments [5]. To measure topic differences between two groups of users, we first compute the posterior probability of each topic separately for the adapted and non-adapted users. We then compute the rate of increase for each topic as the difference between the posterior of the topic using the LLR measure, which is the difference between logarithms of the ratio of posterior probability of the topic in adapted group to non-adapted group.

4.4.3 Visual Content Analysis

With the advent of social networking services, users are increasingly involved in multiple social networks to benefits from diversity of services. For instance, more

[4] To compute LLR measure, we assume that all n-grams are probable in both cohorts with a very low prior probability of $p = 10^{-6}$.

[5] We tuned the number if topics by perplexity as suggested by (Wallach et al., 2009).

than half of US adults (52%) and majority of youngsters (71%) participate in two or more social media sites[6]. Recent studies revealed that users expose different characteristics and behaviours across multiple social networks (Song et al., 2016, Song et al., 2015a). Indeed, analyzing users from different social views provides a better way to comprehensively understand user's behaviours. Hence, we also investigate the differences between AC and NC groups according to the visual contents shared in Instagram social service. Comparing visual concepts of shared images however is a challenging task due to the richness and complexity of shared images. To effectively represent visual contents, we represent each image with a bag-of-visual-concept in which each images is represented with a vector of visual concepts happening in it. Inspired by prior studies (De Choudhury, Sharma, and Kıcıman, 2016, Deng et al., 2009, Farseev et al., 2015), we utilized 1000 visual concepts of ImageNet as a predefined visual concept dictionary due to its popularity in multimedia studies (Deng et al., 2009). We hence constructed a feature vector for each image based on the state-of-the-art deep learning architecture of GoogleNet (Szegedy et al., 2015). We next compute the user's feature vector by averaging the feature vector of all images which were shared by the user.

[6]According to Pew Research Center, Social Media Update, 2014

4.5 Results

4.5.1 LIWC Analysis

Table 4.3 summarizes the LIWC measures of behavioural attributes for the two cohorts of users: adapted and non-adapted users. Overall, the contents published by AC users are less negative than those published by NC users, demonstrating that AC users have a positive perspective towards their health and lifestyle.

Affective attributes. As can be seen from the table, adapted users are less negative. Previous studies also reported similar correlation where negative affection is associated with poor health conditions and engagement (Schwartz et al., 2016). Further, NC contents demonstrate more anger and sadness rather than AC contents. This result may attribute to the fact that being unsuccessful to cope with their issues make patient to be more angry and feel hopeless, loneliness and restless. The impact can be amplified in reverse direction where feeling hopeless and loneliness is highly correlated with less engagement and success, which needs further investigations.

Cognitive attributes. In terms of cognitive attributes, AC patients use more negation structures such as 'not', 'no' as compared to NC patients. Further, they also share perception words, i.e. 'see' and 'feel', which shows they are more likely to express their feeling. Meanwhile, NC users use less certainty in their publishing which is associated to more self-consciousness rather than users who

are able to control their health condition (Taylor and Brown, 1988). This finding is interesting which indicates NC users feel guilty regarding their situation and hence may engage less with their community (See section 4.5.2 for more results).

Linguistic style attributes. NC users have higher lexical density. They also share longer sentences as compared to AC. This is to be expected as NC users utilize social media as a means for acquiring information about their health concerns, as pointed by (De Choudhury, Morris, and White, 2014,Gray et al., 2005); such contents are mostly about self and hence people try to completely describe their situation (Cao et al., 2011). This result needs to be studied more carefully as some studies associate lower lexical density to negative emotions as NC already has shown such characteristics. NC contents are more concerning about past and less focused about future, while AC users are more discussing about future. This is likely attributed to the anxiety of users and their concerns about their health conditions and issues. The literature has leveraged that lower future concern is a known attribute of negative attitude towards user's own life, arising from their problems in managing their health condition(Chapman, Perry, and Strine, 2005). Further, NC users show less social concerns since negative thought are associated with the self. Hence, they are less likely to talk about social concerns and community topics. More surprisingly, AC users are less concern about health and death as compared with NC users. This can own to the fact that these users have already adapted their lifestyle to their situation and health condition, concerning less about their health and their disease consequences. Further, AC users

Table 4.3: The result of Mann-Whitney U-test between posts published by AC and NC according to different behavioural attributes. Each value shows the percentage of words in tweet messages shared by users in each linguistic or psychological category. We used non-parametric test to compute significance.

Category	**AC**	**NC**	**p-value**
Affective			
Positive	4.278	4.168	0.103
Negative	1.580	1.659	0.060
Anxiety	0.269	0.271	0.241
Anger	0.437	0.485	0.004
Sad	0.387	0.412	0.195
Swear	0.138	0.167	0.002
Cognitive			
Negation	1.289	1.315	0.174
Certainty	1.032	1.031	0.247
Cognition	0.757	0.750	0.231
Perception	0.478	0.466	0.222
Linguistic style: Lexical density			
Word Counts	31638	38749	0.002
Word Per Sentence	21	156	0.103
Verbs	10.565	10.669	0.242
Adjectives	3.581	3.635	0.162
Adverbs	2.1×10^{-5}	1.0×10^{-5}	0.191
Linguistic style: Tense			
Past tense	1.925	1.938	0.342
Present tense	7.179	7.308	0.151
Future tense	0.736	0.686	0.040
Linguistic style: Interpersonal awareness			
1st person singular	3.767	3.968	0.096
1st person plural	0.560	0.562	0.046
2nd person	1.427	1.535	0.011
3th person	0.653	0.728	0.014
Linguistic style: Quantities			
quantities+numbers	6.072	5.481	0.009
Linguistic style: Social concerns			
social	5.903	6.194	0.007
family	0.250	0.274	0.037
friend	0.143	0.159	0.121
health	1.900	1.957	0.340
death	0.104	0.131	0.003
work	1.833	1.678	0.056

Table 4.4: The result of n-gram study between posts published by AC and NC.

N-gram	*LLR*	N-gram	*LLR*	N-gram	*LLR*	N-gram	*LLR*
N-grams (AC > NC)				**N-grams (NC > AC)**			
miles hour	7.222	mins felt good	7.179	feeling support	−0.419	num hr later	−0.419
ran miles	6.348	strides	6.348	units novorapid	−0.419	novorapid 2hr later	−0.418
felt good	1.953	hills	1.873	sugar level	−0.418	continued ride cyclemeter	−0.418
keeping	1.873	ride	1.801	suggestion	−0.418	glucose level	−0.418
check strava	1.448	min walk	1.284	hate	−0.418	shouldnt hurt bgnow	−0.418
awesome	1.251	finish	0.887	hurt bgnow	−0.418	high	−0.418
sweatbetes	0.873	beautiful	0.738	really want	−0.415	latest level	−0.389
a sweet life	0.782	cure	0.732	weird	−0.388	crazy sugar	−0.388
ready	0.715	lovely	0.677	stupid	−0.364	shit	−0.361
mysugar	0.642	easy	0.630	how to	−0.331	nightmare	−0.301

have already shown positive affection in their behavior so they may less discuss about negative concepts like death. Last but not least, AC users are more talk about quantities and numbers; this is an important finding and specific to our study. It clearly demonstrates that those are successful to manage their condition are concern about the quantities, which is deeply related to self-management of their condition. This demonstrates that diabetes management demands a careful consideration to balance the lifestyle; adjust their consuming calories, specifically carbohydrates, which will result into a successful management. This finding was already reported by research efforts in health sciences. However, our results reveal that people who discuss about quantities in social media are more likely to follow the correct management program, probably due to the fact that discussing about topic shows its importance for users. Overall our study attests the potential value of social media as a sensor for understanding users' success which can be utilized as an information source in designing better intervention programs and social services as well as investigating public health issues.

4.5.2 Topical Content Analysis

In this section, we investigate how the contents published by these two cohorts of users are different from each other in terms of topics discussed. As mentioned in section 4.4.2, we studied the frequency of different n-grams and topics in posts published by users.

N-grams comparison. We observed a great distinction between the usage of n-grams in posting of AC and NC users. Generally speaking, the content published by two groups of diabetic users demonstrates that AC users mainly act as a social supporter or content providers for diabetic users. This finding itself is interesting and demonstrates that, by designing appropriate intervention tools/programs, we can help diabetic users to better manage their health condition.

Table 4.4 shows a list of 40 different n-grams organized in two different groups with their associated LLR values. The right group lists n-grams with highest LLR values, demonstrating those are important for AC users and the left group summarizes the list of important n-grams for NC users. Overall, our finding verifies the results observed in the last section, as we observe positive n-grams in the first group compared to the last group, which may attribute to the fact that NCs are struggling with their diabetes and focusing to find a proper way to manage their condition. In contrast, AC users are optimistic to the situation and spread positive emotions and experiences. The following contextual themes were observed from the data. (1) We found clear evidences of anxiety and anger (e.g., 'crazy sugar',

'shit', and 'hate') in NC contents owing to the fact that managing diabetes is problematic, in some sense, for this cohorts. This shows that online environment may be a place for them to release their emotional pressure through interacting with their peers. (2) Contents on seeking help and assistance ('how to', 'really want', and 'suggestion') is also evident in NC users. This finding align with the pervious one which shows social platform may be perceived as a supporting environment for patients with diabetes, where not only users seek emotional support but also ask for informational support (Wang, Kraut, and Levine, 2012). Retrospective studies have reported that receiving emotional support is one of the main intentions that attracts users to utilize social networks for health, especially for chronic diseases like diabetes, insomnia, depression and so on (Greene et al., 2011, Jamison-Powell et al., 2012, Taylor and Brown, 1988). (3) AC users, however, more frequently use positive words ('felt good', 'beautiful', 'nice', and 'lovely'). The use of positive n-grams shows a positive view on the life and the tendency of spreading positive emotions and feelings. (4) Compared to NC users, AC users use diabetes management tools and platforms like 'mysugar', demonstrating they are more curious and ambitious on managing their diabetes. Despite the importance and value of using computational framework in managing health problems, the benefits and impacts of using computational frameworks, from simple recording to high-end supporting framework like 'mysugar'[7], and 'onedrop'[8], in managing diabetes is still not fully investigated and more research needs to be investigated.

[7] `https://mysugr.com/`
[8] `http://onedrop.today/`

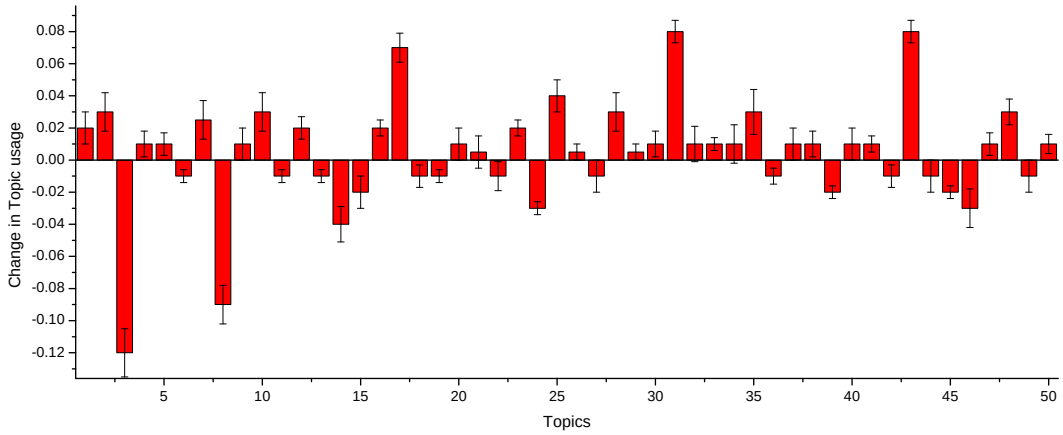


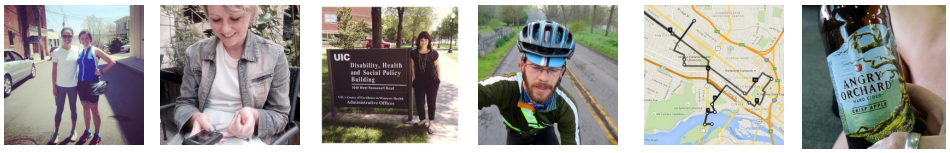
Figure 4.1: Changes in topic usage for two cohorts of diabetes patients, where positive value shows that the topic is important for AC users, otherwise NC users.

Table 4.5: Examples of topics and corresponding representative words

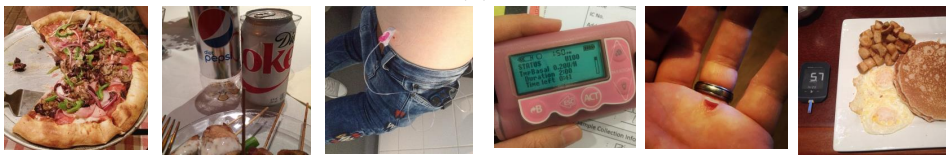
ID	Representative Words	Topic
T3	tired, hate, missed, hurt, horrible, struggled, sick, damn, ugh, lol	Self-critical
T8	afraid, want, useless, comfortable, bad, pain, except, tough, nothing, easy	Conflicting feeling and emotions
T17	running, gym, daily, hypo, mile, ride, walking, sugar, cyclometer, check	Activities and Sports
T31	talk, dsma, advice, bed, insulin, diet, nutrition, sugar, ask, help	Social support
T43	research, interested, diet, DiabeticDiary, prove, fact, fitness, food, sleep, hyperglycemia	Authorized information

Topic analysis. We also extracted the underlying topics exists in the corpus and investigated to which extend the content published by AC and NC users are different from each other. We adopted LDA framework as described in Section 4.4.2 to extract textual topics from document corpus. Figure 4.1 depicts the differences of the existence of topics across two groups. As you can see from the Figure, the mean change across two groups is 20%, which shows two cohorts of users are discussing on different topics online. Specifically, we observe that topics #3 (Self-critical), #8 (Conflicting feelings), #17 (Activities and Sports), #31 (Social support), #43 (Authorized information) show notable variations between two groups. To have a better sense of these topics we listed top 10 representative words for each topic in Table 4.5. From the Figure 4.1, following points can be observed. (1) Topic #3 represents 'self-critical' contents and thoughts about being guilty and negative attitude about self in NC cohort. This can be attributed to the failure of users in self-managing diabetes and similarly their desire to handle the situation. This is also consistent with the literatures, reporting that chronic disease sufferers develops tendencies of self-criticism which sometimes goes beyond the normal level and may result in mental disorders (Anderson et al., 2001, Katon and Sullivan, 1990). (2) Topic #8 represents struggling and conflicting feelings and emotions that are often perceived by NC users. Investigating on the springs and outcomes of these conflicting emotions is worthwhile, which may result to establish better intervention in lifestyle medicine. (3) Discussing about activities and sports is common in AC communities as expressed by topic #43. Indeed a

detail checking of extracted topics shows that two other topics which thematically talks about activities (#25, #28); however, they did not show strong distinction between two groups of diabetes. This result verifies the findings from health sciences which states the positive correlation between regular sport activities like 'running' and better management of diabetes, especially diabetes Type II(Klein et al., 2004). It is worth noting that users discuss about sports and exercise activities in social media often utilize tracking devices linked with web portal and mobile applications, which assist them recording the history of their activities and planning for future. The finding is aligned with a recent research reporting that persistent usage of mobile applications significantly increases the success of users in weight loss program (Park et al., 2015). Despite the increasing popularity of tracker devices and mobile applications, limited studies have investigated their impacts and roles in managing chronic diseases, especially diabetes. (4) Topic #31 describes contents related to social support in online communities. Indeed, #Bgnow acts as a support groups, or a fast-response support group, for patient with diabetes, where they seek and provide informational and emotional support from their peers. This was already verified by the difference between n-grams usage within two groups. (5) Topic #17 reflects authorized information about diabetes, medications and management programs, showing that diabetic patients in AC groups spread more authorized information about diabetes. In essence, professional health providers leverage the power of social media to disseminate health content for health seekers and AC users republish this information in their



(a)



(b)

Figure 4.2: Example images have been shared by AC users (**a**), and NC users (**b**).

network. This result highlights other aspects of social media in healthcare.

4.5.3 Visual Content Analysis

Figure 4.2 shows some representative examples from the images which have been shared by two groups of users. Our analysis demonstrates that the visual contents in shared images of users' Instagram accounts are highly related to the success of users in managing diabetes. Figure 4.3 depicts top 20 statistically significant correlations between visual concepts and the category of users. Several interesting signals can be observed from the Figure. For example, the visual concepts 'mountain bike' and 'unicycle' are positively correlated to AC category which demonstrates the strong preference of AC users to manage their diabetes with lifestyle change. Some other concepts like 'sunglasses', 'crash helmet', and 'street sign', which are objects related to activities, also demonstrate the same preferences. Prior research efforts have also reported similar results in obesity and fitness related studies from social media platforms (Mejova et al., 2015). Conversely, visual concepts 'menu',

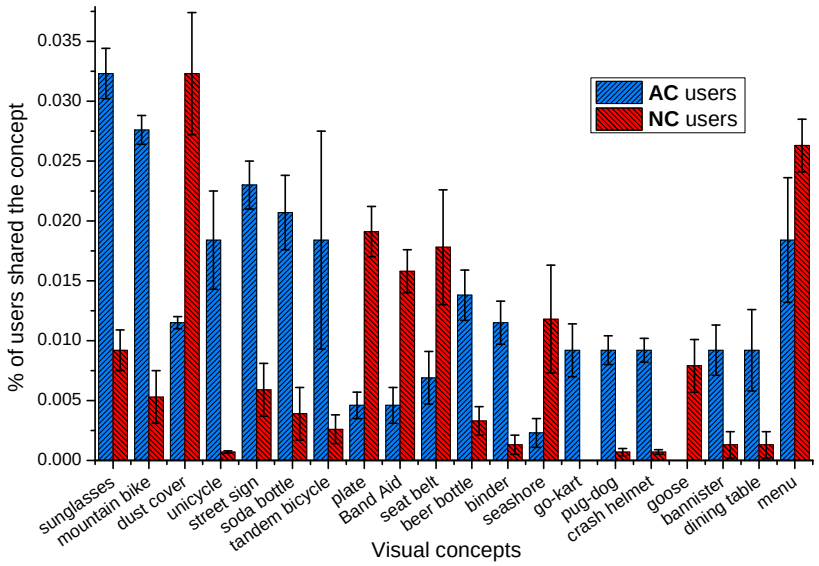


Figure 4.3: Visual concepts which commonly shared by two cohorts of users: AC, and NC users.

and 'plate' are correlated to NC users, which illustrates the possible reason for failing to control their blood glucose level. Further, the positive correlation between the visual concept 'Band Aid' and NC users may indicate that they like to share the picture of their injection site, e.g. Insulin injection or pump site, which shows their anxiety related to their health condition.

4.6 Discussion

Our findings reveal several characteristics of social media, specifically Twitter, for diabetes. Many of our finding align with prior studies, stating that social media is a rich platform for health consumers through which they seek and share health information. Overall, the study revealed that people online behaviours expose their health conditions and states as well as their success in adapting their life style

to their wellness condition. Waving together these observations, it demonstrates that patient generated wellness data on social media can be effectively utilized for inferring users' success in managing his health condition, which can be effectively used for designing better intervention programs and services to assist patient in better management of their diabetes.

4.6.1 Clinical Relevance

The abundant amount of available data can also assist us to better understand patient behaviours and detect potential issues resulting failures in self-management of diabetes. From a clinical perspective, social media can be utilized to complement patient self-report diaries by implicitly tracking his/her online behaviour. Social media can also assist to provide intervention through non-intrusive assessment of content providing and publishing, as discussed below.

Persuasion Oriented Intervention. With a proper lifestyle and behavioural change, we can successfully manage several chronic diseases such as diabetes, and obesity. While it seems an easy task, in practice, changing lifestyle is a challenging and complex task. According the Fogg Behavioural Model (FBM), three elements need to converge together in order to a behaviour occur: motivation, ability, and trigger (Fogg and Hreha, 2010). Indeed, when a behaviour does not occur, one of these elements is missing. Chronic disease sufferers usually have enough motivation to perform the target behaviour, which is suggested by the management program; however, they frequently will not trigger to perform the task on the

correct time, e.g. reduce their sugar consumption or their sedentary lifestyle. By utilizing social networks not only we can understand the user lifestyle and wellness condition, but, more importantly, it is possible to motivate the user and trigger him in the same time. For example, we can suggest him some interesting outdoor activity based on his past preferences or suggest him to have a healthier meal. Further, providing useful health information regarding his/her health condition can effectively motivate user to follow the disease management program, in our case of the diabetes management program.

Social Influence Intervention. In psychology, social influence theory attests that individual's emotions, opinions, and behaviours are affected by others. Social influence has been studied in different domains and environments such as sales, marketing, leadership, and so on (Kelman, 1958). The holistic concept of wellness traditionally has been studied from different aspects includes, physical, mental, social, and spiritual components. Late studies extend this perspective to the social interactions finding that social interactions may affect individual's wellness either in positive and negative manner. For example, recent studies have revealed that person's circle of friends may influence his/her weight(Shoham et al., 2012) and his/her sport activity level, i.e., how active he/she is in sports. Upon these findings, we can assist NC diabetic patients to better manage their health condition through connecting them to AC users, i.e. diabetic users who already find how to successfully manage their disease.

4.7 Related Work

The emergence of social media delineated a shift among users from passive consumption of information to active creation and sharing of contents, which provides new resources to understand and analyse population behaviour. Web search is an important and ubiquitous way through which users acquire information about health conditions and major events like epidemics. Therefore, analysing the query logs of events provides an implicit way to understand public health. Eysenbach (2006) reported the high correlation between click rates on influenza topics and the influenza-linked illness cases reported by CDC. Another study by Yahoo! and CDC reported that the rate of search queries pertaining to cancer is correlated to mortality rate and news coverage of cancer (Chunara, Andrews, and Brownstein, 2012). Google Flue [9] trend is a successful service which predicts flue infection trends based on online queries.

The emergence of Twitter also opened new opportunities to public health researchers to understand the wellness of society. For example, Twitter data was employed for outbreak surveillance of swine flu (H1N1) (Ritterman, Osborne, and Klein, 2009), understanding misinformation on epidemic events, e.g. Ebola (Kalyanam et al., 2015), behavioural change of new mothers (De Choudhury, Counts, and Horvitz, 2013a, De Choudhury et al., 2014), and so on.

Thanks for the richness and veracity of social media data, which provides an

[9] `http://www.google.org/fluetrends`

invaluable opportunity to public health researchers; it is now possible to study public health and wellness in large-scale, which was not possible with traditional methods both in terms of time and cost. While all these research are useful, many of them only focus on public health trends and behaviours, failing to provide any insight about a specific user wellness. Consequently, studying and developing approaches which are able to provide insights to individual's lifestyle and wellness are highly demanded. Towards this research avenue, in this Chapter, we utilize social media to study factors affecting the success of diabetes patients in managing their blood glucose values.

4.8 Summary

Social media is continually being used as a platform for informational and emotional support around health challenges transforming these platforms as a source for knowledge, support and engagement for patients living with chronic diseases such as diabetes. In such a context patients are encouraged to share the exact values of their health measurements such as blood glucose level. In this chapter, we investigated the behavioural distinction of two groups of diabetes patient based on their published posts online. In particular, we studied the behavioural distinction between patients who can successfully manage their blood glucose value and those who fail. We have observed several distinctions in terms of linguistic, textual and visual contents of published posts online. We also provided a supervised approach to predict the success of users based on their online behaviours.

CHAPTER 5

Wellness Representation of Users

Due to the pervasiveness of social media platforms, everyday, millions of users increasingly utilize social networks such as Twitter and Instagram to share their wellness data and to fulfil their health demands. Effective mining of patient generated wellness data (PGWD) can provide actionable insights into the wellness of individuals as well as collaborative behaviour of communities. While data-driven approaches are increasingly used for personalized healthcare (He et al., 2014a, Liu et al., 2015, Xu, Sun, and Bi, 2015), as an important and distinct data source, understanding PGWD available on social networks presents great opportunities to improve care delivery.

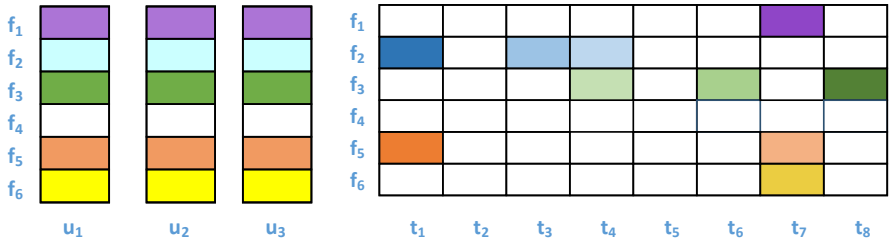
Representation learning, also called latent feature learning or in general feature learning, has become an effective tool for many machine learning and data mining applications and is still an open problem (Jin et al., 2015, Weston, Weiss, and Yee, 2013, Zhao, McAuley, and King, 2014, Zhao, McAuley, and King, 2015). The hypothesis behind representation learning is to find a low-dimensional embed-

ding of data instances while preserving different discriminative factors of variation behind the data. Regarding the importance of data representation, we propose to explore representation learning approach towards analyzing and understanding users' wellness from PGWD. In particular, in this chapter, we demonstrate a representation learning approach to learn the latent profile of users from their social media contents.

5.1 Challenges

Despite its value and significance, PGWD in social networks has not been fully utilized due to the following challenges. (1) *Longitudinality.* Wellness data are longitudinal per se, which means multiple measurements or repeated events are available for each subject (Liu et al., 2015,Xu, Sun, and Bi, 2015,Zhou et al., 2014). For example, Hemoglobin A1c (HbA1c) test might be done several times per year for diabetic patients. The longitudinal nature of the problem provides a matrix of wellness data describing patient at different time points (Wang et al., 2013, Xu, Sun, and Bi, 2015,Zhou et al., 2014). This is quite different from standard machine learning representation where we have a static vector of features, as shown in Figure 5.1. In such a context, time dimension plays an essential role. (2) *Noisiness and Incompleteness.* As perviously discussed (see Chapter 1 and Chapter 3), social media is a highly varied and informal media; arising from various background and intention of users (Wang et al., 2011). Moreover, missing data is an intrinsic nature of PGWD since patients do not persistently report their wellness data. In most

cases, users are not keen enough to expose the event or they self-censor the content due to privacy concerns (De Choudhury, Morris, and White, 2014,Lin et al., 2014). This means that the absence of a wellness event in PGWD does not always mean that the event did not happen. (3) *Heterogeneity.* An intrinsic characteristic of the wellness domain is heterogeneity of the patient population according to their health conditions; meaning that wellness attributes and events related to each user can be highly different from the others (Nori et al., 2015). For instance, even though diabetic users often share similar characteristics, they are still different from each other based on demographic attributes (e.g., age and gender), type of disease (e.g., Type I Diabetes, Type II Diabetes, Gestational Diabetes, etc.), and many other behavioral and genetic factors. Even though patient stratification is a well-established approach in health informatics (Wang, Zhou, and Hu, 2014), this kind of disease-specific context has not been fully investigated in many wellness models such as re-admission prediction (He et al., 2014a), disease progression modelling (Wang, Sontag, and Wang, 2014, Zhou et al., 2013), risk prediction (Wang et al., 2014); and the assumption of a homogenous cohort does not hold in the population. How to share information among homogenous population while simultaneously avoid interactions between heterogeneous populations is still an open problem in wellness modelling.



(a) (b)

Figure 5.1: Vector-based and Longitudinal representation, where different colors show distinct features and color intensity shows relative value of the feature. (a) Representation of three distinct users in vector-based approach; vector-based approach represents a single measurement for each feature; (b) Representation of one user in longitudinal approach with 8 different time points. Longitudinal data represents each feature with a set of values pertaining to different time points.

5.2 Overview

To deal with the challenges raised by the distinct PGWD, in this chapter, we investigate to learn wellness representation of users from social media. Our framework, in contrast to conventional models, determines the wellness latent space directly from users' longitudinal data, instead of attribute-value data, by considering two types of domain priors, namely the heterogeneity in data space and temporal contingency of wellness concepts. In particular, the proposed approach decomposes longitudinal data into two components: wellness latent space, and temporal representation of users. To effectively handle data heterogeneity, the learned wellness latent space is comprised of two sub-spaces, i.e., shared and personalized latent spaces, as shown in Figure 5.2. The learned temporal representation is constrained to model the temporal progression of wellness attributes and simultaneously tackle the problems arising from missing data values. The proposed framework has been extensively examined through several machine learning tasks to evaluate its effec-

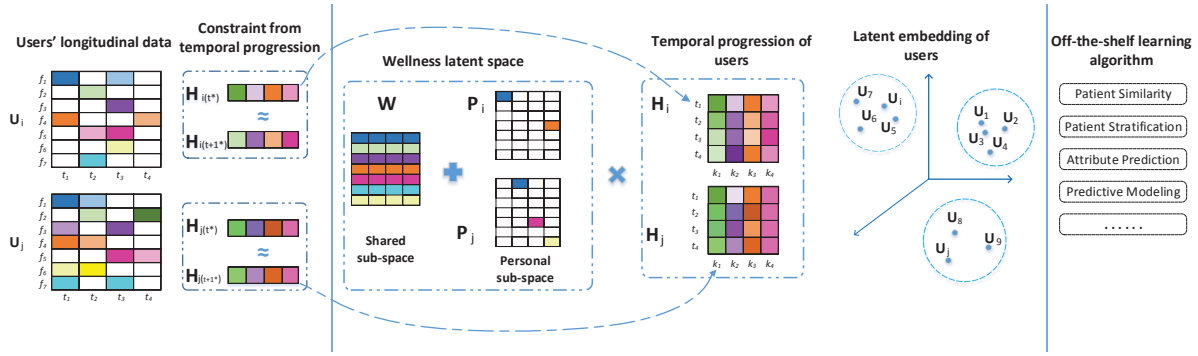


Figure 5.2: The conceptual view of the proposed framework for representation learning of longitudinal data from social networks. The wellness latent space is comprised of two sub-spaces: shared and personal latent space. The final representation of each user, i.e., $\mathbf{H}_i$, embeds the user in the latent space while each row is his/her representation at one time point, where different colors show distinct features and color intensity shows relative weight of the feature.

tiveness in user embedding.

The main advantageous of the proposed representation learning are as follows:

- We propose a representation learning approach for longitudinal wellness data available in social networks. Specifically, we decompose longitudinal PGWD into wellness latent space and the temporal progression of users in that space.

- We exploit consistency within homogenous population as well as distinction between heterogeneous population to learn a shared and personalized latent space for embedding users.

- We incorporate the temporal progression prior of wellness data in the learning process to tackle the problems arising from missing and sparsity of data.

- We propose an efficient approach to find the embedding of the users in the two sub-spaces which scales well. This is an important feature permitting us to use the framework in web scale.

5.3 Problem Statement

In this section, we first present the notations and then formally define the problem of representation learning of longitudinal data. Note that the problem we study is different from traditional representation learning since the latter merely focuses on learning a latent representation for "flat" attributed-value data, in contrast to longitudinal data; only projecting high-dimensional vectors to a low-dimensional space.

5.3.1 Problem Formulation

Let $\mathcal{U} = \{\mathbf{U}_1, \mathbf{U}_2, \dots \mathbf{U}_n\}$ denote a set of n users' longitudinal information. Each user longitudinal information $\mathbf{U}_i$ is denoted by $\mathbf{U}_i \in \mathbb{R}^{f \times t}$, where f is the number of different wellness events and features[1] and t is the length of observation window in which we measure the events. Note that the user's longitudinal data is a matrix where $\mathbf{U}_{i(j,k)}$ represents the measurement value of the wellness event j at time point k for the user i.

We want to learn a low-rank representation of users in $\mathcal{U}$ so that if two users u and v have similar wellness data, their representation would be closer. We assume that the longitudinal data can be factorized to two components: a latent space representing wellness concepts and the temporal progression of each user in the latent space, as shown in Figure 5.2. The factorization process is capable

[1]In this text, we use wellness feature (e.g., blood glucose, hypertension) and wellness events (onset of asthma attack, hyperglycemia) interchangeably.

of reconstructing the user data matrix on observed values. In general, a user's longitudinal representation is formally defined as a matrix $\mathbf{H}_i$, where each row of the matrix, i.e., $\mathbf{H}_{i(j*)}$, represents the user wellness state at time point j.

With the notation above, we formally define the longitudinal user representation problem as: *Given a set of users' longitudinal information* $\mathcal{U}$*, we aim to learn a model as follows,*

$$f : \mathcal{U} \rightarrow \{\mathbf{W}_i, \mathbf{H}_i\}, \tag{5.1}$$

which can compute ***wellness latent space*** $\mathbf{W}_i \in \mathbb{R}^{f \times k}$ *and* ***temporal progression*** *of each user in the wellness latent space, i.e.,* $\mathbf{H}_i \in \mathbb{R}^{t \times k}$*.*

The final representation of each user, i.e., $\mathbf{H}_i$, precisely embeds the user in wellness latent space while each row is his/her representation at one time point.

5.4 Factorization of Longitudinal Data

As mentioned, PGWD includes two major aspects: wellness aspect and temporal aspect. Constructing an effective representation requires to subtly decompose these two components from each other. The key hypothesis behind longitudinal data factorization is that user's data matrix can be decomposed into two factors: (1) wellness latent space, and (2) the temporal onset of wellness events over observation windows, i.e., time dimension.

5.4.1 Preliminaries

Retrospective studies have shown that the wellness features can be projected to a latent space with a lower dimensionality; resulting in a dense representation of the original features (Zhou et al., 2014). This factorization process is capable of reconstructing the observed entries of original matrix, i.e., patient longitudinal wellness data. Inspired by these research findings, we utilized nonnegative matrix factorization (NMF) to decompose patient data matrix into two low rank matrices which are capable of approximately reconstructing the observed matrix. NMF is a matrix factorization algorithm that factorizes the non-negative data matrix into two positive matrices (Lee and Seung, 2001). Assume that $\mathbf{U}_i \in \mathbb{R}^{f \times t}$ represents the data matrix for patient i, the aim of factorization is to decompose $\mathbf{U}_i$ into to non-negative matrices $\mathbf{W}_i \in \mathbb{R}^{f \times k}$ and $\mathbf{H}_i \in \mathbb{R}^{t \times k}$, whose product provide a good approximation of $\mathbf{U}_i$, i.e., $\mathbf{U}_i \approx \mathbf{W}_i \mathbf{H}_i^T$, where k is a pre-specified parameter denoting the dimension of reduced space. For instance, in topic modeling, k represents the number of topics while it denotes the number of desired latent dimensions in feature learning. Formally, NMF aims to minimize the following objective function,

$$\min_{\mathbf{W}_i, \mathbf{H}_i} \|\mathbf{U}_i - \mathbf{W}_i \mathbf{H}_i^T\|_F^2 \quad \text{s.t.} \quad \mathbf{W}_i \geq 0, \mathbf{H}_i \geq 0, \tag{5.2}$$

where $\mathbf{W}_i$ is called the *wellness basis matrix* and $\mathbf{H}_i$ is the *temporal progression matrix*. Intuitively, $\mathbf{H}_i$ represents how wellness dimensions evolve over time for

the given user. In other words, it demonstrates how the user's wellness is going to improve, stable, or worsen as time passes. As the above objective function is not jointly convex in $\mathbf{W}_i$ and $\mathbf{H}_i$, finding the global minima is infeasible (Lee and Seung, 2001). Therefore, alternating minimization is iteratively utilized to find a local minima. The iterative update rules are as follows,

$$\mathbf{W}_i \leftarrow \mathbf{W}_i \odot \frac{\mathbf{U}_i\mathbf{H}_i}{\mathbf{W}_i\mathbf{H}_i^T\mathbf{H}_i}, \quad \mathbf{H}_i \leftarrow \mathbf{H}_i \odot \frac{\mathbf{U}_i^T\mathbf{W}_i}{\mathbf{H}_i\mathbf{W}_i^T\mathbf{W}_i}. \tag{5.3}$$

where $\odot$ and the division symbol in this matrix context denote element-wise multiplication and division, respectively. Note that the above setting is different from standard matrix factorization where $\mathbf{U}_i$ represents an item-feature matrix constructed from the whole dataset.

It is also worth noting that there are several useful properties in using matrix factorization (Gu, Zhou, and Ding, 2010, Tang et al., 2013) for sub-space learning: (1) the non-negative property of NMF ensures an intuitive decomposition of the patient matrix into wellness and temporal parts, in contrast to other matrix factorizations that do not hold this property, e.g., PCA and SVD; (2) the model has a nice probabilistic interpretation with Gaussian noise; (3) many existing optimization approaches can be utilized to find an optimal solution for the the model; (4) it can be scaled to a large number of users, which is a common setting in social media platforms; (5) this formulation is flexible and allows us to introduce prior knowledge such as heterogeneity and temporality of the wellness attributes.

5.4.2 Shared Wellness Space for Homogenous Cohort

Factorization of user's longitudinal data provides an intuitive decomposition of data matrix of a given user into wellness latent features and their temporal progression over time. However, decomposing wellness data of each user in isolation may not provide effective representation due to the excessive sparsity in data. Besides, comparing latent spaces of different users would be a challenging task since the factorization process may extract diverse latent features fitted on each user data. Therefore, extracting a common latent space from the entire collection of data is normally preferred. The hypothesis behind collective latent space learning is that the wellness latent space extracted from different data instances, in our case users, should admit the same underlying structure, corresponding to higher-level latent features constructed from the combination of lower level features. At the same time, the temporal progression of these wellness latent features can vary from user to user depending on user's attributes, behaviors and so on. Mathematically, it can be formulated as the following objective function,

$$\begin{aligned}\min_{\mathbf{W},\mathbf{H}_i} J_{SLS} = \frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{U}_i - \mathbf{W}\mathbf{H}_i^T\|_F^2 + \frac{\lambda_1}{2}(\|\mathbf{W}\|_F^2 \\ + \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{H}_i\|_F^2) \quad \text{s.t.} \quad \mathbf{W} \geq 0, \mathbf{H}_i \geq 0,\end{aligned} \tag{5.4}$$

where the first term factorizes users' longitudinal data, while the second and third terms control the complexity of models. Here, $\mathbf{W}$ is to compute the shared wellness

latent space among all patients. The above objective function assumes that all patients share the same wellness space and learns a unique mapping **W** from the original feature space to the wellness latent space.With sharing of the latent space among all patients, we indeed transfer knowledge among the patient cohorts, which is attractive especially when the available information for each patient is limited and the cohort is homogenous (Pan et al., 2014, Song et al., 2015a). Sharing also reduces the effect of noise since the latent space is derived from a large amount of data.

5.4.3 Personalized Wellness Space for Heterogeneous Cohort

Even though learning a common latent space from dataset is an intuitive and well-established tradition in machine learning, its performance is highly varied in real applications since it assumes a rigid consensus in dataset; i.e., all the data instances need to follow a specific latent space (Pan et al., 2014). This is, however, impossible in real situations since patients can be divided into different cohorts with different characteristics. For example, diabetic users can be divided into three major patient groups: type I, type II, and gestational diabetics and several minor groups merely based on disease type, where each group holds different characteristics (Groop, 2015, Nori et al., 2015, Sun, Wang, and Hu, 2015). This suggests that we need a personalized feature learning framework to deal with heterogeneity in data space.

Inspired by the notion of "dirty models" in machine learning for handling heterogeneous high-dimensional data (Jalali et al., 2010, Jin et al., 2015), we assume that individual's wellness latent space can be slightly deviated from the shared space extracted from the whole population. Mathematically, we consider the following learning model,

$$\begin{aligned}\min_{\mathbf{W},\mathbf{H}_i,\mathbf{P}_i} J_{PLS} = &\frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{U}_i - (\mathbf{W}+\mathbf{P}_i)\mathbf{H}_i^T\|_F^2 \\ &+ \frac{\lambda_1}{2}(\|\mathbf{W}\|_F^2 + \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{H}_i\|_F^2) + \frac{\lambda_2}{n}\sum_{i=1}^{n}\|\mathbf{P}_i\|_1 \\ &\text{s.t.} \quad \mathbf{W}\geq 0, \mathbf{H}_i \geq 0, \mathbf{P}_i \geq 0, \end{aligned} \tag{5.5}$$

where the latent space is estimated by the summation of two parameters $\mathbf{W}$ and $\mathbf{P}_i$. The first part of Eq. (5.5) learns three sets of parameters: (1) $\mathbf{W}$ is the shared latent space for all users inferred from the entire dataset; (2) $\mathbf{P}_i$ is to model heterogeneity in the data space, i.e., the personalized feature space; and (3) $\mathbf{H}_i$ demonstrates the temporal evolution of each individual in the latent space. By imposing different regularizations for each parameter, we can fit an effective personalized learning model. The above formulation includes two set of regularizers; the second term, i.e., $(\|\mathbf{W}\|_F^2 + \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{H}_i\|_F^2)$, controls the generalization performance of the model to avoid overfitting and the third term (ℓ_1-norm) leads to a sparse model. It is worth noting that the aforementioned model extends the concept of dirty model to longitudinal data (Jalali et al., 2010).

From clinical aspects, the proposed model is closely related to precision medicine (Groop,

2015, Mirnezami, Nicholson, and Darzi, 2012), where medical treatments are tailored to individual patients based on their detailed genetic and clinical profiles as well as lifestyle factors. By learning personalized latent space, i.e. $\mathbf{P}_i$, our model follows precision medicine paradigm through modeling distinct characteristics of individuals. Our model also considers disease principle paradigm by providing a computational model with the shared feature space, i.e., $\mathbf{W}$, where disease treatment and prevention are learned from the entire population. This also presents significance in treating patients with missing values.

5.4.4 Modeling Temporal Information

Recall that wellness attributes smoothly evolve over time. The temporal progression of wellness attributes suggests that these values gradually changes over time (Liu et al., 2015, Xu, Sun, and Bi, 2015). Thus, modelling the temporal evolution of wellness attributes can effectively reduce the noise and sparsity of the wellness data through imputation of missing values as pointed by (Sun, Wang, and Hu, 2015, Xu, Sun, and Bi, 2015). As each row of the temporal progression matrix $\mathbf{H}_{i(j*)}$ indicates the wellness representation of the user i at time point j, we hence penalize the sudden changes of wellness attributes between neighbouring time points. Specifically, the temporal progression of wellness attributes can be mathematically modelled as,

$$\mathcal{R}_{temporal} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{t-1} \|\mathbf{H}_{i(j*)} - \mathbf{H}_{i(j+1*)}\|^2, \tag{5.6}$$

where $\mathbf{H}_{i(j*)}$ denotes the wellness representation of the user i at time point j. To facilitate the optimization of the temporal progression term, Eq.(5.6) can be restated in an equivalent form as follows,

$$\mathcal{R}_{temporal} = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{t-1}\|\mathbf{H}_{i(j*)} - \mathbf{H}_{i(j+1*)}\|^2 = \frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{H}_i\mathbf{R}_i\|_F^2, \tag{5.7}$$

where $\mathbf{R}_i \in \mathbb{R}^{t\times t-1}$ is the temporal smoothness indicator and is precalculated by the following definition,

$$\mathbf{R}_{i(j,k)} = \begin{cases} 1 & \text{if } j = k; \\ -1 & \text{if } j = k+1 \\ 0 & \text{otherwise .} \end{cases} \tag{5.8}$$

Intuitively, Eq.(5.7) imposes that the wellness representation of the given user at two consecutive time points be close to each other.

5.5 Algorithm Details

The optimization framework, which integrates prior information into representation, is defined as follows,

$$J_{Space} + \alpha\mathcal{R}_{temporal}, \tag{5.9}$$

where the first term, i.e., J_{Space}, denotes the objective function for learning latent space, i.e. Eq.(5.4) and Eq.(5.5) for homogenous and heterogenous settings, respectively. Meanwhile, the second term incorporates temporal prior of wellness attributes into the learning model.

In this section, we introduce an efficient algorithm to solve the optimization problems and discuss its time complexity. Note that the optimization problem of homogeneous setting is a special case of the heterogenous setting. Therefore, we only provide the algorithm for heterogenous setting. Here, by substituting Eq.(5.5) in the above equation, we have the following cost function,

$$\begin{aligned}\min_{\mathbf{W},\mathbf{H}_i,\mathbf{P}_i} \mathcal{O} = & \frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{U}_i - (\mathbf{W}+\mathbf{P}_i)\mathbf{H}_i^T\|_F^2 \\ & + \frac{\alpha}{2n}\sum_{i=1}^{n}\|\mathbf{H}_i\mathbf{R}_i\|_F^2 + \frac{\lambda_1}{2}(\|\mathbf{W}\|_F^2 + \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{H}_i\|_F^2) \\ & + \frac{\lambda_2}{n}\sum_{i=1}^{n}\|\mathbf{P}_i\|_1 \\ & \text{s.t.} \quad \mathbf{W} \geq 0, \mathbf{H}_i \geq 0, \mathbf{P}_i \geq 0, \end{aligned} \tag{5.10}$$

where α, λ_1, and λ_2 are regularizers to control the trade-off between different components.

5.5.1 Optimization Algorithm

We adopt an alternating optimization strategy to find the optimal values for model parameters. Specifically, we alternatively update $\mathbf{W}$, $\mathbf{H}_i$, and $\mathbf{P}_i$ to minimize the objective function while keeping the others fixed. To enforce the non-negativity

constraints, we need to incorporate Lagrange multipliers. Let Λ_w, Λ_{pi}, and Λ_{hi} be the Lagrange matrices for constraints $\mathbf{W} \geq 0$, $\mathbf{P}_i \geq 0$, and $\mathbf{H}_i \geq 0$, respectively. The Lagrange $\mathcal{L}$ is:

$$\mathcal{L} = \mathcal{O} + Tr(\Lambda_w \mathbf{W}) + \sum_{i=1}^{n} (Tr(\Lambda_{pi} \mathbf{P}_i) + Tr(\Lambda_{hi} \mathbf{H}_i)). \tag{5.11}$$

Optimizing W

By fixing $\mathbf{H}_i$ and $\mathbf{P}_i$, we can rewrite the objective function as follows,

$$\begin{aligned} \min_{\mathbf{W}} \mathcal{L} = \frac{1}{2n} \sum_{i=1}^{n} \|\mathbf{U}_i - (\mathbf{W} + \mathbf{P}_i)\mathbf{H}_i^T\|_F^2 + \frac{\lambda_1}{2} \|\mathbf{W}\|_F^2 \\ + Tr(\Lambda_w \mathbf{W}) + C, \end{aligned} \tag{5.12}$$

where C is constant with respect to $\mathbf{W}$. Taking the derivative with respect to $\mathbf{W}$, we have,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{P}_i \mathbf{H}_i^T \mathbf{H}_i - \mathbf{U}_i \mathbf{H}_i) + \frac{1}{n} \sum_{i=1}^{n} \mathbf{W} \mathbf{H}_i^T \mathbf{H}_i + \lambda_1 \mathbf{W} + \Lambda_w. \tag{5.13}$$

Using the Karush-Kuhn-Tucker (KKT) complementary condition, we have the following update rule for $\mathbf{W}$,

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\sum_{i=1}^{n} \mathbf{U}_i \mathbf{H}_i - \sum_{i=1}^{n} \mathbf{P}_i \mathbf{H}_i \mathbf{H}_i^T}{\sum_{i=1}^{n} \mathbf{W} \mathbf{H}_i^T \mathbf{H}_i + n\lambda_1 \mathbf{W}}. \tag{5.14}$$

Optimizing $\mathbf{P}_i$

By ignoring terms that are independent of $\mathbf{P}_i$ in Eq.(5.11), the objective function boils down to:

$$\min_{\mathbf{P}_i} \mathcal{L} = \frac{1}{2n}\|\mathbf{U}_i - (\mathbf{W} + \mathbf{P}_i)\mathbf{H}_i^T\|_F^2 + \frac{\lambda_2}{n}\|\mathbf{P}_i\|_1 + Tr(\Lambda_{pi}\mathbf{P}_i). \tag{5.15}$$

The above objective function is non-smooth since it is the composition of a smooth term and a non-smooth term, i.e., ℓ_1 penalty, and gradient descent method is not available for solving the formulation. Inspired by (Chen et al., 2009, Nesterov, 2004), we utilize the accelerated proximal method (APM) to solve its equivalent smooth reformulation. APM has been excessively utilized in data mining and machine learning communities (Chen et al., 2009,Hu et al., 2013) due to its optimal convergence rate among all first-order techniques and its ability of dealing with large-scale non-smooth optimization problems. Note that we focus on discussing the key concepts of APM, i.e, the proximal operator and its efficient computation; the detailed description of APM can be found in (Nesterov, 2004).

APM maintains two sequences of variables: a feasible solution sequence $\{\mathbf{P}_i^j\}$ and a searching point sequence $\{\mathbf{S}^j\}$, where the superscript, i.e., j, shows the index in the sequence. We denote the smooth and non-smooth part of the objective function $\mathcal{L}$ by $f(.)$ and $g(.)$. APM reformulates the optimization problem by a

proximal operator which is formally defined as,

$$\mathbf{P}_i^{j+1} = \arg\min_{\mathbf{P}_i^j} \mathcal{M}_{\gamma^j, \mathbf{S}^j}(\mathbf{P}_i^j), \tag{5.16}$$

where,

$$\mathcal{M}_{\gamma^j, \mathbf{S}^j}(\mathbf{P}_i) = f(\mathbf{S}^j) + \langle \nabla f(\mathbf{S}^j), \mathbf{P}_i^j - \mathbf{S}^j \rangle + \frac{\gamma^j}{2} \|\mathbf{P}_i^j - \mathbf{S}^j\|_F^2, \tag{5.17}$$

where $\mathbf{S}^j$ is computed based on the past solutions by $\mathbf{S}^j = \mathbf{P}_i^j + \tau^j(\mathbf{P}_i^j - \mathbf{P}_i^{j-1})$ and $\nabla f(\mathbf{S}^j)$ denotes the derivatives of the smooth component $f(.)$ in the objective function, i.e., Eq.(5.15), at the search point $\mathbf{S}^j$. The parameter γ^j is the step size and is determined by line search according to Armijo-Goldstein rule. By ignoring terms that are independent of $\mathbf{P}_i^j$ the objective function boils down to:

$$\mathbf{P}_i^{j+1} = \arg\min_{\mathbf{P}_i^j} \|\mathbf{P}_i^j - \mathbf{Q}^j\|_F^2, \tag{5.18}$$

where $\mathbf{Q}^j = \mathbf{S}^j - \frac{1}{\gamma^j} \nabla f(\mathbf{S}^j)$ and indeed the solution of $\mathbf{P}_i^j$ is the Euclidian projection of $\mathbf{Q}^j$ onto convex set of constraints (Nesterov, 2004). Here, $\nabla f(\mathbf{S}^j)$ denotes the gradient of the smooth component $f(.)$ in Eq.(5.15) at $\mathbf{S}^j$, which is defined as:

$$\nabla f(\mathbf{P}_i) = \frac{1}{n}(\mathbf{W}\mathbf{H}_i^T\mathbf{H}_i + \mathbf{P}_i\mathbf{H}_i^T\mathbf{H}_i - \mathbf{U}_i\mathbf{H}_i). \tag{5.19}$$

Optimizing $\mathbf{H}_i$

To minimize the cost function with respect to $\mathbf{H}_i$, we first fix $\mathbf{W}$ and $\mathbf{P}_i$, and then compute the derivative with respect to $\mathbf{H}_i$ as follows,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{H}_i} &= \frac{1}{n}[-\mathbf{U}_i\mathbf{P}_i - \mathbf{U}_i^T\mathbf{W} + \mathbf{H}_i(\mathbf{W}+\mathbf{P}_i)^T(\mathbf{W}+\mathbf{P}_i)] \\ &\quad + (\frac{\lambda_1}{n}\mathbf{I} + \frac{\alpha}{n}\mathbf{R}_i\mathbf{R}_i^T)\mathbf{H}_i, \end{aligned} \tag{5.20}$$

where $\mathbf{I}$ denotes the identity matrix with correct dimensions. Using the Karush-Kuhn-Tucker (KKT) complementary condition, we have the following update rule for $\mathbf{H}_i$,

$$\mathbf{H}_i \leftarrow \mathbf{H}_i \odot \frac{\mathbf{U}_i^T\mathbf{P}_i + \mathbf{U}_i^T\mathbf{W}}{\mathbf{H}_i(\mathbf{W}+\mathbf{P}_i)^T(\mathbf{W}+\mathbf{P}_i) + (\lambda_1\mathbf{I} + \alpha\mathbf{R}_i\mathbf{R}_i^T)\mathbf{H}_i}. \tag{5.21}$$

It is worth noting that the convergence of the updating roles can be proven using standard auxiliary function approach introduced in (Lee and Seung, 2001).

5.5.2 Computational Complexity and Convergence

We now analyze the time complexity of our learning framework using big O notation. The learning algorithm includes three main steps for optimizing three set of variables, i.e. $\mathbf{W}$, $\mathbf{P}_i$, and $\mathbf{H}_i$. In update rule for $\mathbf{W}$, the time complexity is $O(nkft)$, where n is the number of users, k is the dimension of latent space, f is the dimension of original feature space, and t is the length of the observation

window. The main computational time for $\mathbf{P}_i$ is to compute the derivation of smooth part of objective function, i.e., Eq.(5.19), which is $O(ftk)$. As we need to update $\mathbf{P}_i$ for all samples, in our case each user, the total computational time is in order of $O(nkft)$. The computation for $\mathbf{H}_i$ is similar to $\mathbf{P}_i$ with time complexity of $O(nkft)$. If we need q iteration for updating the values of variables, the time complexity of the final algorithm is in order of $O(qnkft)$. As t denotes the length of observation window and it is in the size of few hundred, which is a small constant, in our experiment it is a six months period and $t = 25$, the final complexity can be approximated by $O(qnkft) \approx O(qnkf)$, making PLS a linear representation learning algorithm. We empirically verified this in our experiments, as the actual running time of our framework was similar to running plain NMF on all longitudinal data matrices.

Note that the complexity of time dimension is less critical, as discussed, because that in most cases, the time dimension of the patients are often less than 1000. Recall that the finest time unit of the longitudinal data of users is day. Using weekly granularity, 1000 time dimension covers up to 20 years of records.

Considering the convergence aspect, the optimization function is non-convex in with respect of all three variables in the Eq.(5.10). However, it is convex with respect of each variable. Hence we can use coordinate descent approach to find the minimum of the function. Coordinate descent in the function would converge to a stationary point and it would be a local minimum for the function. Please note that the function is non-smooth with respect to $\mathbf{P}_i$. To solve the optimization

problem with respect to $\mathbf{P}_i$, we utilized the proximal approach which projects the non-smooth part of the problem to a convex constraint. (Liu, Ji, and Ye, 2009b) demonstrated that the convergence rate of the optimization approach is $O(\frac{1}{\sqrt{\epsilon}})$ where ϵ is the desired accuracy.

5.6 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of the proposed representation learning of users from social networks in both homogenous and heterogeneous settings. We used our approach in two real-world datasets to accomplish different tasks, which show superiority of our proposed approach over the state-of-the-art baseline methods.

5.6.1 Experimental Settings

Datasets

Diabetes Dataset. We evaluated our approaches on a real-world dataset containing posting of diabetic users about diabetes and their associated symptoms, medications, and activities. To construct the dataset, we first gathered a set of users who actively utilized diabetes related hashtags like "#diabetes" and "#bgnow" or follow diabetes support groups, such as American Diabetes Association, in Twitter microblogging service. Table 5.1 shows the list of hashtags and twitter support groups which were used for collecting candidate twitter users.

Table 5.1: The list of seed hashtags and twitter support group used for collecting twitter user pool.

Hashtags		**Support Groups**	
#Dibetes	#Bgnow	@AmDiabetesAssn	@WDD
#Diabetic	#T1D	@DiabeticConnect	@DiabetesUK
#type2diabetes	#T2D	@diabetesdaily	@NDEP
#diabeteschat	#Doc	@DiabetesMine	@citiesdiabetes
#LivingwithDiabetes	#Dblog	@DiabetesHealth	@diabeteshf

We next crawled the twitter profile of these users using Twitter API and selected the users who explicitly mention diabetes as an interest in their Twitter profile, resulting into 14,108 different candidate user accounts. To construct ground truth labels, we utilized an automatic approach, inspired by similar efforts in computational social science (Lin et al., 2014), based on users who self-declared their disease information. We used expressions like "I am (Type—T) (1|2) diabetic" to extract disease type for each user based on his/her profile information [2]. Disease type here refers to the major types of diabetes and includes three categories: Type I diabetes, Type II diabetes, and Others. We merged all the other non-common diabetes types as one category[3]. Table 5.2 shows the statistics of our dataset. As you can see, we could extract the health attributes of more than 50 percent of users (7,474 Twitter accounts) based on their self-declared information in their profiles, which we will use for the evaluation of our framework. Table 5.3 shows some example profiles from our collected dataset and their associated regular expression

[2]We followed a bootstrapping approach similar to (Thelen and Riloff, 2002) to ensure the coverage and diversity of used patterns, where all extracted patterns are manually verified to ensure accuracy.

[3]In our dataset, there are three non-common diabetes types: gestational diabetes, diabetes LADA (Type 1.5), and diabetes insipidus.

Table 5.2: Statistics of the Diabetes Dataset

# of Users		14,108
# of Tweets		11,491,036
Disease Type	Diabetes Type I	4,194
	Diabetes Type II	2,477
	Others	803

Table 5.3: Example profiles from our diabetes dataset

Husband. Dad. I've diagnosed as Type 1 diabetic since DATE. On a journey ...	I *diagnose* Type (1—2) diabetic	Type 1
I LOVE LIFE!! I am type 2 diabetic and take insulin	I * Type (1—2) diabetic	Type 2
Writer, avid reader, ...; live with T1 diabetes, ...	* with (T1—T2) diabetes	Type 1

and ground truth labels[4].

BG Dataset. This is the dataset which was constructed in Chapter 3 and also used in Chapter 4. The dataset comprises of Twitter activities of diabetes patients who actively share their wellness information on Twitter. They not only post about their lifestyle information and activities such as their diet, activities, and emotional states but also share their health information in terms of medical events and measurements like their blood glucose value, HbA1c test results and hypoglycaemia/hyperglycaemia onset. In Chapter 4, we labelled all users in the dataset with "successful", and "unsuccessful" tags showing that he managed to maintain an on-target blood glucose value or failed to do so, respectively. We used this dataset to evaluate the effectiveness of our method in predicting the wellness states of users (such as the blood glucose value) based on the longitudinal wellness data of users on social media. This is important since wellness states are highly

[4]Due to user privacy concerns, some words/sentences may be different from original version.

Table 5.4: Statistics of the BG dataset

# of Users	1,174
# of Tweets	1,060,105
# Successful Users	436
# Unsuccessful Users	738

dependent on historical values, i.e. temporally dependent, showing that we need to consider longitudinal information of user's wellness instead of merely considering current state. Table 5.4 shows the statistics of this dataset.

Extraction of Longitudinal Wellness Descriptions

Feature extraction is an important aspect in our approach since it determines the original representation of data. However, compared with textual documents in traditional media, a distinct characteristic of texts in social media platforms is that they are noisy and short (Tang et al., 2012). To comprehensively represent user's wellness, inspired by studies in clinical text mining (Akbari et al., 2016, Aronson, 2001, Xu et al., 2010), we extracted three kinds of features as follows.

1) *RxNorm description*. Medication information is one of the most important types of wellness data. It is critical for healthcare safety and quality as well as for prognostic modeling (Zhou et al., 2006). Extracting medication information from free text reports is a traditional but challenging problem in clinical text processing (Doan et al., 2012, Chhieng et al., 2006, Sohn et al., 2014, Zhou et al., 2007). To extract medication information, we utilized the approach proposed in (Xu et al., 2010) which utilizes semantic parser and domain knowledge to accurately extract medication information, i.e. medication names and signatures, from

free texts and was commonly used as medication representation in literature.

2) *UMLS description*. We also used a widely-used knowledge-based system called MetaMap to assign Unified Medical Language System (UMLS) Meta-thesaurus semantic concepts to user's social posts (Aronson, 2001). MetaMap is a rule based system that assigns UMLS Meta-thesaurus semantic concepts to phrases in natural language text. MetaMap is commonly used as a complementary resource containing tremendous amount of medical knowledge, which is independent from training dataset, in contrast to other systems. We collected all MetaMap's finding in the dataset and used their gold standard medical concepts as features. Along with the analogy of bag-of-words, we constructed a Bag-of-Concepts (BoC) in medical terminology and represent each user in the resulting space. The final BoC contains $5,370$ distinct concepts.

3) *Personal Wellness Events*. Personal wellness events are defined as events that are directly related to wellness of individuals; providing a summary of users' lifestyle and wellness such as diet event, medication use, and hospitalization. Patients frequently post these events in their social accounts. We utilized the approach proposed in Chapter 3 to extract personal wellness events from users' published messages on Twitter. This will provide a high level description of user's wellness state; containing 14 distinct dimensions.

To construct the longitudinal wellness matrices of users, we utilized social media posts of users. We need to select a granularity level in time dimension and extract the information according to the selected granularity. We observed that

the daily granularity is too sparse with more than 0.95% of users seemed reluctant to report information daily. We thus constructed the users' longitudinal data at the weekly granularity. As we collected the data for six months, from May to October 2015 and constructed 25 time points for the entire period [5].

Evaluation Tasks and Metrics

To demonstrate the effectiveness of the proposed representation learning approach, we implicitly evaluated its performance in two commonly-used machine learning settings: supervised and unsupervised learning. The hypothesis behind implicit evaluation is that a good representation will improve the performance of the selected tasks as compared to other baselines. We hence evaluated our problem in two supervised problems: attribute prediction and success prediction and one unsupervised problem community detection, where communities were extracted by clustering of users in the user latent space.

Attribute Prediction. Attribute detection was widely applied in user profiling to infer latent attributes of users such as age and gender prediction, education and occupation detection, political party detection (Farseev et al., 2015, Gottipati et al., 2013). As inferring wellness attributes is a critical step in many downstream applications like recommendation (Wing and Yang, 2014), we hence proposed to predict wellness attributes of users using information from social media. We evaluated the performance of learning representation in predicting disease type which

[5]We did not consider the first week of May and the last week of October because the data was partially crawled.

is the major wellness attribute of users. To evaluate our approach, we utilized diabetes dataset with 10-fold cross validation and reported the performance in terms of precision, recall, and the area under the receiver operating characteristic curve (AUC). Due to the imbalance nature of the dataset, the latter provides a good explanation of the effectiveness of the proposed method (Powers, 2011).

Success Prediction. Success prediction is the task of predicting whether a specific user can successfully maintain his/her health indicators in a suggested range. For example, a diabetic patient who can successfully control his blood glucose value in the healthy range would be categorized as a successful patient, otherwise an unsuccessful patient. Due to its importance in wellness domain (Weber and Achananuparp, 2015), we evaluated our feature learning framework in predicting users' success in managing their diabetes, i.e., maintaining their blood glucose value in the healthy range. Here, we considered the success prediction as a binary classification problem and utilized **BG** dataset to evaluate our problem.

Clustering. We also evaluated our representation learning approach under the clustering task. Compared to classification, clustering is totally unsupervised and heavily relies on the learned features and similarity measure. We adopted the commonly used cosine similarity for clustering of users in the learned latent space. We compared the performance of different approaches in terms of accuracy and normalized mutual information (NMI) on diabetes dataset.

5.6.2 On Performance Comparison

To the best of our knowledge, we are the first to study feature learning of the longitudinal data in social media. To demonstrate the effectiveness of representation learning approaches, we compared our learned features with those of other state-of-the-art unsupervised feature learning methods, while keeping the classification and clustering scheme fixed. We compared the following baseline methods:

- **ALL**. All original features are adopted for each user.
- **LapScore**. Laplacian score evaluates feature importance by its ability to preserve the local manifold structure of data (He, Cai, and Niyogi, 2005).
- **Spec**. Features are selected by spectral analysis. This approach can be considered as an extension of Laplacian score method (Zhao and Liu, 2007).
- **NDFS**. Nonnegative discriminate unsupervised feature selection via joint nonnegative spectral analysis and $\ell_{2,1}$-norm regularization (Li et al., 2012b).
- **Shared Latent Space (SLS)**. Users are embedded into shared latent space of Eq.(5.4).
- **Personal Latent Space (PLS)**. Each user's is represented using personalized latent space learned from Eq.(5.5) which models both temporality and heterogeneity.

We followed previous research studies to tune the parameters for all baseline methods (He, Cai, and Niyogi, 2005, Li et al., 2012b). The neighborhood size has

been fixed to 5 for **LapScore** and **NDFS**, as suggested to be the best in (He, Cai, and Niyogi, 2005,Li et al., 2012b). There are some regularization parameters for **NDFS**, and **LapScore**, which were set based on the experiments from the original papers. **SLS**, and **PLS** have three different regularizer parameters α, λ_1, and λ_2. In the experiments, we empirically set $\alpha = 0.1$, $\lambda_1 = 10$, and $\lambda_2 = 0.4$ using grid search and 10-fold cross validation. More details about the effects of these parameters on the proposed framework will be discussed in Section 5.6.3 and 5.6.4.

We evaluated the predictive performance of the proposed framework in supervised setting using attribute prediction and success prediction experiments. From the learnt feature space, we derived features by averaging the latent features along the time dimension within a given observation window (25 weeks). The performance of attribute prediction and success prediction is presented in Table 5.5 in terms of precision, recall, and AUC. From the Table, we can observe the following points: (1) Feature selection is important as well as effective. The selected features not only can reduce the computational time of the algorithm (Zhao and Liu, 2007) but more importantly can improve the final prediction performance, where all the feature learning approaches outperform **ALL** baseline. (2) **LapScore** and **Spec** have a neck to neck performance with a slight improvement by **LapScore** which is consistent with the results reported in past research efforts (Li et al., 2012b, Zhao and Liu, 2007). (3) **NDFS** often outperforms both **LapScore** and **Spec** which is attributed to the feature selection process in **NDFS**. **LapScore**

and **Spec** analyze features individually which may overlook possible correlation between distinct features, as reported in (Li et al., 2012b), while **NDFS** considers feature correlation. (4) **SLS** and **PLS** consistently outperform other baseline methods on both tasks. For example, **PLS** approximately gained up to 6% and 3% relative improvement in terms of precision in attribute prediction and success prediction, respectively. The reason is probably that **SLS** and **PLS** takes advantages of temporal correlation between feature values to mitigate problems arising from data sparsity and missing values. However, all baseline methods assume the *i.i.d* assumption, which is not valid in the wellness domain (Xu, Sun, and Bi, 2015). Moreover, **PLS** outperforms **SLS** most of the time, which shows the importance of modeling heterogeneity in data space, as reported by past efforts (Jin et al., 2015, Liu et al., 2015). Overall, these observations support the fact that joint learning features and modeling domain prior knowledge would achieve the best performance (Liu et al., 2015, Sun, Wang, and Hu, 2015).

We also evaluated our method under unsupervised setting, i.e., clustering. Table 5.6 summarizes the result of clustering users in learned latent space in terms of accuracy and NMI. The results are similar to that for supervised setting, i.e., classification. (1) **SLS** and **PLS** approaches outperform all the baseline methods in terms of accuracy and NMI, which demonstrates the importance of modeling temporal progression of wellness features as well as feature learning. The reason is probably that vector-based representation cannot capture the context around each user probably due to excessive sparsity of data, noisy information in social

Table 5.5: Performance of attribute and success prediction

Disease Type Prediction						
	All	**LapScore**	**Spec**	**NDFS**	**SLS**	**PLS**
Prec	42.31	44.71	41.50	46.32	53.02	59.34
Recall	42.66	46.11	44.82	43.71	48.21	54.20
AUC	63.05	64.47	62.35	67.33	69.85	72.15
Success Prediction						
Prec	62.21	67.34	64.08	68.82	71.33	74.12
Recall	67.45	66.72	64.31	65.01	68.20	68.75
AUC	64.10	61.20	61.40	68.95	72.21	76.80

Table 5.6: Performance of users clustering

	All	**LapScore**	**Spec**	**NDFS**	**SLS**	**PLS**
ACC	51.32	56.10	52.84	54.88	56.11	58.01
NMI	0.0224	0.0227	0.0233	0.0240	0.0272	0.0287

media, and inability to model temporal evolution of user. (2) **PLS** can effectively improve the performance with relative improvement of 2% over **SLS**, in terms of accuracy. This improvement is attributed to the effectiveness of modeling heterogeneity of the patient populations, i.e., different sub-populations in patients, which is modeled in **PLS** while **SLS** assumes a homogeneous cohort of patients. Overall, the proposed method of *joint modeling temporality of wellness features and heterogeneity of user space* can outperform other baselines and achieve the state of the art performance. This result is consistent with several past research in multi-feature machine learning where dirty models are used to model heterogeneity in samples (Liu et al., 2015, Sun, Wang, and Hu, 2015).

5.6.3 On the Effect of Temporal Information

We are now interested in figuring out the effectiveness of different components in our proposed model. In particular, we compared the performance of incorporating temporal smoothness of wellness features in our model. i.e., $\mathcal{R}_{temporal}$. We hence conducted experiments to comparatively validate the following experimental settings:

- **PLS**. Our proposed framework which models both heterogeneity and temporality, i.e., Eq.(5.5).
- **SLS**. Our proposed framework which models temporality with homogenous assumption, i.e., Eq.(5.4).
- **PLS-noTP**. We did not consider the temporal smoothness in **PLS** by setting $\alpha = 0$.
- **SLS-noTP**. We did not consider the temporal smoothness in **SLS** by setting $\alpha = 0$.

We only reported the results for the success prediction task since similar observations have been made for the other tasks. The results of component-wise analysis are reported in Table 5.7. From the table, the following observations can be made: (1) **SLS-noTP** achieves the worst results. This can be explained by the fact that **SLS-noTP** neither models the temporal smoothness in wellness features, nor considers the heterogeneity in the patient population. These results

Table 5.7: Effectiveness evaluation of each involved component in our proposed models.

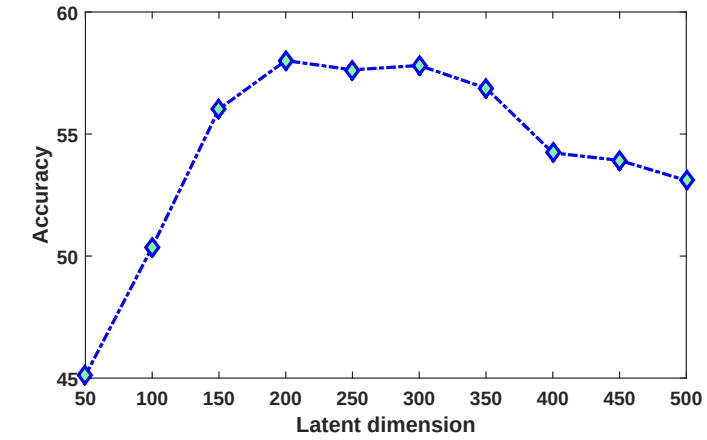
	Precision	Recall	P-value
PLS	74.12	68.75	-
SLS	71.33	68.20	3.1e-3
PLS-noTP	64.02	58.91	1.7e-3
SLS-noTP	62.37	56.09	2.4e-4

imply the importance of joint modeling the temporality of wellness features and heterogeneity of the patient population. (2) **SLS** and **PLS** consistently outperform their counterparts **SLS-noTP** and **PLS-noTP**, which significantly supports the importance of modeling temporality of wellness features. This result has also been reported in modeling disease progression based on patient's EHR (Xu, Sun, and Bi, 2015, Zhou et al., 2006). (3) **PLS** is superior to others; demonstrating that all components in our proposed model is indispensable.

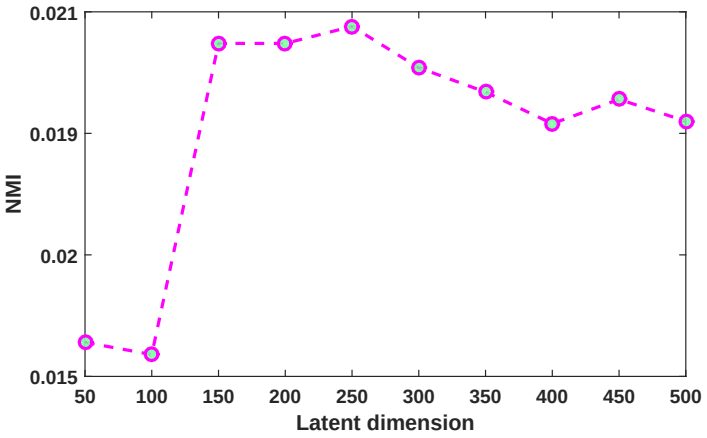
It is worth noting that we also conducted a significance test based on the precision of success perdition task. In particular, we performed paired t-test between our **PLS** model and other baseline methods based on 10-fold cross validation and the results shows that the improvements of our proposed model are statistically significant (p-values are smaller than 0.05).

5.6.4 On Parameter Sensitivity

We also studied the parameter sensitivity of our proposed method. Our model holds two sets of parameters: (1) the latent space dimension, i.e., k; and (2) the regularizers α, λ_1, and λ_2 in Eq.(5.5). We first evaluated the sensitivity of the proposed approach to the dimension of the latent space and then examined



(a)



(b)

Figure 5.3: Effect of latent space dimension. Small values of latent dimension result into limited discrimination power, and large values yield overfitting.

the effects of other parameters in combination with latent space to see how the parameters affect the learned latent space. We only performed parameter study for clustering task to save space.

We first vary dimension of the latent space k in the range of $\{50, 100, 150, \ldots, 500\}$ while fixing the other parameters, i.e., α, λ_1, and λ_2. Figure 5.3 illustrates the clustering performance in terms of accuracy and NMI. The clustering performance is the best when the number of latent dimensions is around 200. The results show

that when the number of latent dimensions is too small, the model is unable to find a good representation. In contrast, a large latent dimension tends to overfit and results in loss of performance. It is worth noting that how to determine the number of features is still an open problem in data mining (Li et al., 2012b).

To assess the effect of parameter λ_1 which controls the complexity of the model, we varied λ_1 as $\{0.001, 0.01, \ldots, 100\}$ while fixing λ_2, and α. Figure 5.4a demonstrates the sensitivity of our framework with respect to various λ_1, and k values. With the increase of λ_1, the clustering performance rises rapidly and then keeps stable between the range of 1 to 10. A high value of λ_1 controls the effects of noise; making the model more robust. The results also demonstrates that the performance is more sensitive to the number of latent dimensions than λ_1.

We studied the effect of parameter λ_2 which controls the personalization aspects of feature learning; making the model more robust in heterogeneous data. Similarly, we changed λ_2 in the range of $\{0.001, 0.01, \ldots, 100\}$ while making other parameters fixed. The results are shown in Figure 5.4b. It can be seen that the performance of our model significantly improved when λ_2 varies between 0.1 and 1, verifying that modeling heterogeneity in the patient population is vital in wellness domain.

We finally investigated the trade-off between temporal smoothness of wellness features and latent space dimension by varying α in $\{0.001, 0.01, \ldots, 100\}$ as presented in Figure 5.4c. As shown in the Figure, in most cases, the clustering performance first increases, reaches its peak and then gradually decreases. The

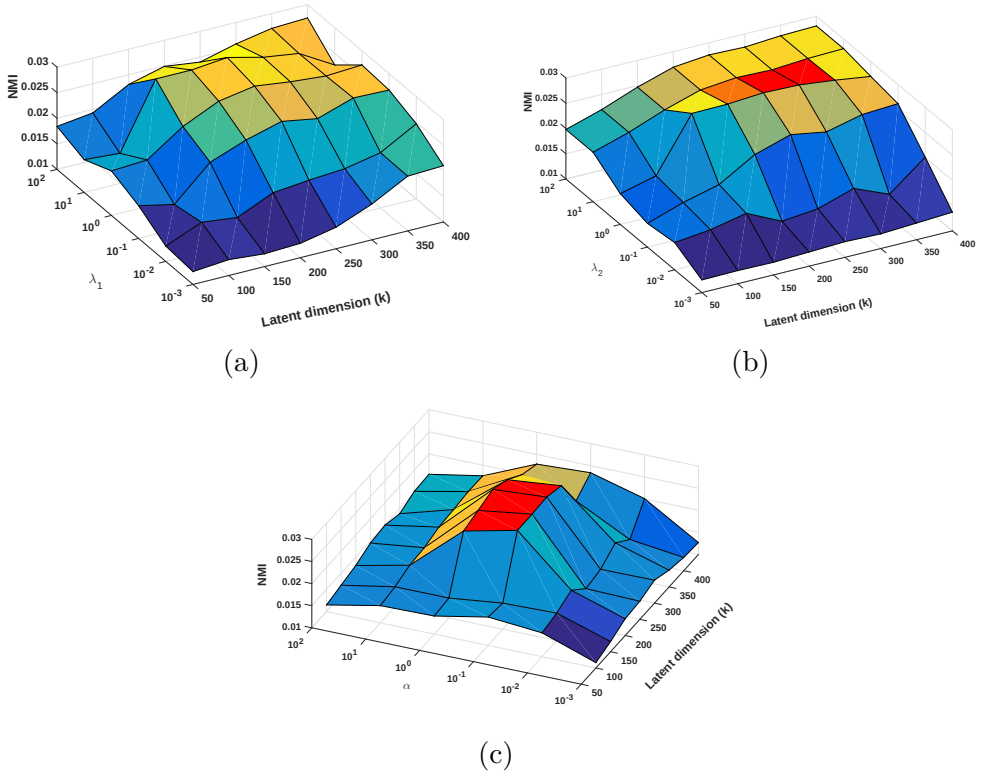


Figure 5.4: The effect of different regularizers on final latent features. Overall, latent dimension is an important factor in learning good representation. Besides, finding the best values for hyperparameters results into learning an effective latent space.

best performance was achieved when α is around 0.1. These observations suggest the importance of modeling both temporal smoothness of wellness features as well as feature learning.

5.7 Related Work

Representation learning, also called latent feature learning, has been widely used as an effective tool for many machine learning and data mining tasks to derive an effective latent space from original data (Jin et al., 2015, Weston, Weiss, and Yee, 2013, Zhao, McAuley, and King, 2014, Zhao, McAuley, and King, 2015). The key idea of representation learning is to seek a low-dimensional embedding of data instances while preserving different discriminative factors of variation behind the data. Recently, factorization based methods have been attracting a lot of interests in modeling user behaviors and interests due to its ability to alleviate data sparsity (He et al., 2014b, Jin et al., 2015, Zhao, McAuley, and King, 2014, Zhao, McAuley, and King, 2015). For example, MaxMF (Weston, Weiss, and Yee, 2013) was developed to represent each user with a set of latent factors representing his/her different latent interests. Zhao et al. (Zhao, McAuley, and King, 2014) incorporated social connections into latent space to improve the performance of recommendation. Seen from the personalization aspect, Zhao et al. (Zhao, McAuley, and King, 2015) proposed a personalized feature projection method that employs users' projection matrices and items' factors to solve one-class recommendation problem.

While various techniques have been proposed for learning representation in machine learning and data mining, most of the existing approaches for latent factor learning have been designed for vector-based representation to embed users (or items) in a low dimensional space. They will fail to provide effective representation if applied to longitudinal wellness data. Furthermore, existing feature learning assumes that data items are *i.i.d.*, which is clearly violated in longitudinal data. Moreover, most of these approaches fail to model heterogeneity in data space or model temporal dependency as a regularized multi-task learning framework but overlook heterogeneity in data space. Our aim is to learn a latent representation directly from longitudinal data where temporality and heterogeneity of data are jointly modeled.

In the area of data-driven health care, phenotyping has been applied to Electronic Health Records (EHRs) to predict the onset of congestive heart failure (CHF) and end stage renal disease (ESRD) by learning a general model (Zhou et al., 2014). Our framework, however, is different from their approach since we simultaneously model the shared latent space between homogenous populations to transfer knowledge among homogenous population as well as learn personalized latent space for each user to learn individual-based features. Their framework either considers a shared space or an individual latent space, which can be considered as a special case of our formulation, i.e., **SLS**. Similarly, Wang et al. (Wang, Zhou, and Hu, 2014) proposed a clustering-based approach to model the heterogeneity in the patient population, where the shared latent space is learnt for each group

of users. It is worth noting that multi-task learning paradigm was also used for investigating EHRs, where they mostly assume the task are homogenous and learn task models simultaneously (Nori et al., 2015, Zhou et al., 2006).

5.8 Summary

In this chapter, we introduced a novel representation learning approach for longitudinal wellness data. The proposed method jointly models the temporal progression of wellness attributes as well as the heterogeneity in the patient populations. In particular, we factorized user's longitudinal data into two components, namely, the latent space representation and user temporal evolution in the space. The latent space is comprised of two sub-spaces: shared latent space and personalized latent space, which permits to exploit both consistency within homogenous cohorts as well as difference amongst heterogeneous cohorts to share an effective representation. Extensive experiments on two real-world datasets and different learning tasks in wellness domain verified the potential ability of the proposed framework in learning a good user embedding.

CHAPTER 6

Discovering and Profiling User Groups and Communities

Social media has been integrated into our life as an inevitable tool for seeking, sharing, and spreading information, opinions, and experiences. The abundance and growing usage of social media along with its ubiquity have significantly changed our information sharing and socialization behaviours. One fundamental task in such network data is to detect salient communities among individuals, aiming at understanding collaborative behaviour of users as well as investigating individuals' behaviour in the context of the group[1]. Hence, discovering user communities, which is the task of finding tightly connected and highly similar user groups, has attracted much attention in recent years. While a large body of work has been devoted to user profiling in social media, little has explored profiling user groups in order to understand the formation of groups as well as construct bet-

[1]In this text, we use community and group, interchangeably.

ter user groups (Harvey, Crestani, and Carman, 2013, Majumder and Shrivastava, 2013, Shin et al., 2015).

6.1 Motivation and Challenges

Group profiling can provide valuable insights about the group and users' behaviors which is important to many social media services. First, group profiling can help understand collective behavior of users and the rationale of group formation. In other words, it explains why individuals join the community. Second, learning the group profile enables us to complete individual's profile in the context of their group affiliations. This is important for handling sparsity and noise of user-level information in social platforms (Li et al., 2015). Besides, understanding social structure underlying users' interactions provides a macro-level understanding of network, compared to micro-level user information, facilitating several applications such as visualization and navigation of networks (Namata et al., 2007), tracking interest shifts of communities (Zhou, Jin, and Liu, 2012), and online marketing (Wang, Guo, and Lan, 2014, Zhao, McAuley, and King, 2015), just to name a few. For instance, recommender system can provide better recommendation by exploiting aggregated interest of the group's members. Further, it facilitates visualization and navigation of large scale networks through coarse and fine grain profiling of communities and sub-communities.

Retrospective studies in social media and data mining have proposed useful approaches for discovering social communities from network and user informa-

tion. For example, modularity decomposition has been applied to link information in social networks (Newman, 2006, Ruan, Fuhry, and Parthasarathy, 2013), and communication pattern has been used for expert team detection (Lappas, Liu, and Terzi, 2009), and generative models were used to find topical communities(Zhou, Jin, and Liu, 2012). While a large body of work has been devoted into community detection, yet many challenges remain to be addressed. First, many existing techniques utilize either network information or content analysis to discover communities (Leskovec, Lang, and Mahoney, 2010, Newman, 2006). However, neither information alone is satisfactory for accurate community affiliation estimation (Yang et al., 2014b). It is attributed to the extremely sparse link information and noisy contents in social platforms, which may significantly drift community discovery process and result in poor performance (Yang et al., 2009). Second, people can perform a wide range of activities in social networks ranging from publishing a post to following their friends to directly communicating through replies. Although exploiting these heterogenous behaviours are essential for discovering good communities, how to integrate them in a unified model is a challenge (Yang et al., 2014b). Third, in many real-word scenarios, some prior knowledge about the community affiliation of users might be available. Take the wellness domain as an example; label information about patients' disease type provide clear evidences about the community affiliation of users, i.e., users with same disease type (e.g., diabetes type I, Type II, etc) belong to the same community. Further, discovering communities without the prior knowledge might

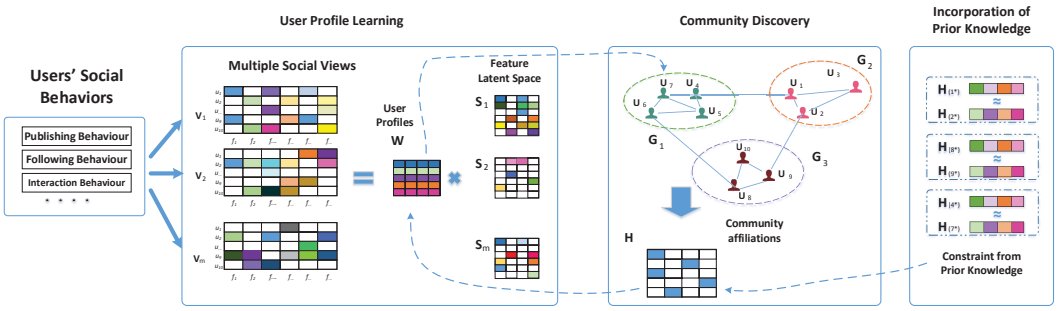


Figure 6.1: The conceptual view of the proposed framework for joint profiling of users and communities in social networks. The framework integrates different social views into a latent space in which we learn the profile of users, **W**, and their community affiliations, **H**. Prior knowledge is incorporated into the discovery process by imposing constraints on the affiliation vectors of the users.

result to ill-posed communities which are not interpretable in real-world scenarios. However, how to incorporate the domain knowledge in the learning and discovery process is a popular and still an open problem. Last but not least, most existing community detection techniques fail to provide any rationale and insight about the formation of the community as well as the collective behavior of the members. How to derive a community profile from its member data pose a great challenge.

6.2 Overview

To tackle these challenges, in this chapter, we propose a learning framework which simultaneously learns the profile of users and communities in social networks, permitting to tackle the challenges mentioned in Section 6.1. As shown in Figure 6.1, in contrast to conventional models, the proposed framework performs community discovery in the latent space, which considers the similarity of users' profiles. In particular, we first exploit different social behaviors of users into various social

views. To effectively handle the sparsity and noise in social media data, we integrate different social views of the network into a low-dimensional latent space representing users' profiles. Next, the optimal community structure is learnt by imposing a similarity constraint over the affiliation vector of users, which seeks dense clusters of users in the latent space. We seamlessly incorporate prior knowledge about the community structure into the community discovery process and turn the process into an optimization problem, where community profile is constructed using a linear pooling operator integrating the profiles of the members. Taking the wellness domain as an example, we constructed a large scale real-world dataset of twitter users who post about diabetes and its associated concepts such as medication use, symptoms, etc. Extensive experiments on the dataset have demonstrated the effectiveness of the proposed approach on discovery and profiling communities as well as leveraged several interesting insights about users' interactions in social media.

The main contributions of this study are as follows,

- To mitigate problems arising from noise and sparsity of data, we seamlessly integrate various social behaviors of users into a unified latent space. Specifically we propose an approach to learn the profile of users and communities from the combination of all social behaviors of users in the social network.
- To learn the optimal embedding of users and their community affiliation, we propose to simultaneously learn the embedding and community affiliation of the users through a semi-supervised approach which is guided by prior

knowledge.

- To learn the community profile, we propose to pool the members' profiles through a simple linear computation which scales well.

6.3 Problem Formulation

The problem we study in this paper is different from traditional community detection approaches since the later normally discovers communities in the feature space by considering either the topology of the network or the contents published by the users. We instead discover communities in a latent space which is constructed from the fusion of different social views in the network. Further, we simultaneously learn user's embedding and community structure, which allows us to directly compare users and groups in the latent space. In this section, we first present the notations and then formally define the problem of joint profiling of users and community in a social networks.

Let $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$ denote a set of n different users in a social network, who can perform m different behaviors $\mathcal{B} = \{b_1, b_2, \ldots, b_m\}$. We construct different social views $\mathcal{V} = \{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_m\}$ based on these behaviours, where the i-th social view, i.e. $\mathbf{V}_i$, represents users based on the i-th social behaviour, b_i. In particular, we construct m social view matrices $\mathbf{V}_i \in \mathbb{R}^{n \times f_i}$, where n, and f_i denote the total number of users and the number of features in the i-th social view, respectively. In general f_i can be any low-level (e.g. words, users) or high-level (e.g.

hashtags, entities) feature. Further, prior knowledge about community structure is available in the form of positive and negative observation pairs $\mathcal{O} = \{\mathcal{O}_+, \mathcal{O}_-\}$, where any pair $(u_i, u_j) \in \mathcal{O}_+$ shows that two users u_i and u_j belong to the same community and $(u_i, u_j) \in \mathcal{O}_-$ indicates that they belong to different communities. Knowing that these users form k different communities, we aim at harvesting their online data to co-profile users and communities as well as estimate the affiliation of users to the discovered communities.

With the above notations, the problem of joint user and community profiling can be formally defined as:

Joint User and Community Profiling: *Given a set of n users on social media, their representation in m different social views $\mathcal{V}$, and prior knowledge $\mathcal{O}$ about community structure, we aim to learn a model as follows,*

$$f : \{\mathcal{U}, \mathcal{V}, \mathcal{O}\} \rightarrow \{\mathbf{W}, \mathbf{G}, \mathbf{H}\}, \tag{6.1}$$

which can compute users' latent profiles $\mathbf{W} \in \mathbb{R}^{n \times l}$, communities' latent profiles $\mathbf{G} \in \mathbb{R}^{k \times l}$, and the optimal community affiliation matrix $\mathbf{H} \in \mathbb{R}^{n \times k}$ such that $\mathbf{H}_{(ij)} = 1$ demonstrates the membership of user u_i to community j, otherwise $\mathbf{H}_{(ij)} = 0$. $\mathbf{W}$ denotes the user profile matrix and $\mathbf{G}$ denotes the community profile matrix, where $\mathbf{W}_{(i*)}$, and $\mathbf{G}_{(j*)}$ represent the latent profile of the i-th user and the j-th community, respectively.

In the rest of this chapter, we introduce an optimization framework which

can jointly learn the profile of users, the optimal community structure, and the profile of communities. We then evaluate the performance of the proposed method through extensive experiments.

6.4 Dataset Description and Representation

In this Section, we first describe dataset construction and ground-truth labeling and next explain how to extract different social views, where each of them represents the users based on one social behavior.

6.4.1 Dataset Description

As previously mentioned, people often utilize social media services for seeking and sharing wellness information. Naturally, they follow and participate on online support groups like “diabeteslife”, “diabetesconnect” to discuss about specific wellness topics. Most popular support groups hold an account in Twitter to publish online content and inteact with their members. Thus we utilized these support groups and their participants for evaluation. To construct the dataset, we first gathered a set of users who followed these diabetes support groups in Twitter. These users have already expressed their interest in joining the community and discussing on diabetes and its related challenges. Table 6.1 shows the list of seed support groups that have been used for selecting diabetes users. We then crawled the twitter profile of these users in November 2015. We removed all the users who have not published any short messages in the last month to filter out those who

Table 6.1: The list of seed support group in Twitter which were used for collecting twitter user pool.

Support group	# users	Support group	# users
@AmDiabetesAssn	1245	@WDD	1120
@DiabeticConnect	1240	@DiabetesUK	940
@diabetesdaily	1210	@NDEP	425
@DiabetesMine	1185	@citiesdiabetes	300
@DiabetesHealth	1145	@diabeteshf	1270

Table 6.2: Summary of different defined social views.

ID	Social aspect	Definition	Relation: Description
1	Network	$u_i \xrightarrow{follow} u_j$	**Following:** u_i follows u_j
2		$u_i \xrightarrow{folllow} u \xleftarrow{follow} u_j$	**Co-Following:** u_i, and u_j follow the same user u
3		$u \xrightarrow{follow} u_i$, and $u \xrightarrow{follow} u_j$	**Co-Followed:** u_i, and u_j are followed by the same user u
4	Interaction	$u_i \xrightarrow{reply} u_j$	**Reply:** u_i replied a tweet from u_j
5		$u_i \xrightarrow{retweet} u_j$	**Retweet:** u_i re-posted u_j tweet
6	Content	$u_i \xrightarrow{post} w_j$	**Content:** u_i's posts containing w_j word
7	Semantic	$u_i \xrightarrow{used} h_j$	**Semantic:** u_i used h_j hashtag

were inactive and failed to participate in the community. This process resulted into 10085 total number of users which we used for the evaluation. We crawled all online behaviours of these users including their tweets, retweets, and reply posts along with their social network to construct the dataset[2].

6.4.2 Data Representation

Users in social media can perform various activities such as content publishing (e.g., post a tweet), network construction (e.g., following a friend), and direct communication (e.g., reply[3] a post). Each of these activities represents the users from different aspects or social views, which together can completely reveal users' behaviours and interest. We propose to build different social views based on

[2]We used Twitter search API and only crawled the latest 3200 tweets for each user due to the twitter API limitation.

[3]In Twitter, a user can reply a post by clicking on the reply icon appear on each tweet.

different social behaviors of the users to comprehensively represent each user. We hence defined six social views which belong to four main categories representing users from different aspects of: *network*, *interaction*, *content*, and *semantic*. Table 6.2 shows different social views of users and we describe how to construct these views in the following sub-sections.

Network-centric Views

In social media research, it is widely accepted that interests and affiliations of users are correlated with that of their social connections and friends, referring as homophily and contagious theory(Shalizi and Thomas, 2011). We hence utilize the network connections between users to represent the social context around them. We quantitatively capture these kind of connections by three types of relations namely "Following", "Co-following" and "Co-followed", as shown in Table 6.2. Inspired by (Hu et al., 2013), we utilize the social connection of the users as social features and construct a user-user matrix for each type of relation. To capture the following relation, we intuitively define $\mathbf{V}_1$ as,

$$\mathbf{V}_{1(ij)} = \begin{cases} 1 & \text{if } u_i \text{ follows } u_j. \\ 0 & \text{otherwise.} \end{cases}, \tag{6.2}$$

where $\mathbf{V}_{1(ij)}$ shows the following relationship amongst users. Similarly, we define two matrices for co-following and co-followed views as the number of shared friends and the number of users who follow both u_i and u_j, respectively. It is worth noting

that the two matrices are normalized by dividing each element by the largest element in the matrix. For example, the co-following matrix is normalized by,

$$\mathbf{V}_{2(ij)} = \frac{\mathbf{V}_{2(ij)}}{\max_{i,j} \mathbf{V}_{2(ij)}} \tag{6.3}$$

As a result, we have generated three views which capture different aspects of connections between users.

Interaction-centric Views

In addition to sharing posts and contents, users in social networks can interact with each other in many different forms such as reply each others statuses, and re-tweeting messages. Interaction in social networks is another important aspect of social communities (Zhao et al., 2015). Retrospective studies have demonstrated that these interactions, while too sparse, provide stronger evidences of similar interest and affiliations (Yang et al., 2014b). We hence define two different interaction views: reply and re-tweets. For each pair of users, we compute the average number of such interactions as their interaction strength. More specifically, the reply interaction matrix can be computed as follows,

$$\mathbf{V}_{4(ij)} = \frac{1}{2}\{|r_{i\to j}| + |r_{j\to i}|\}, \tag{6.4}$$

where $|r_{i\to j}|$ shows the total number of replies from the i-th user to the j-th user. Note that we compute the interactions in two directions since any direction shows

an evidence of connection. In this way, we construct two distinct views $\mathbf{V}_4$, and $\mathbf{V}_5$ from interaction of users and normalize them as we do with network-centric views.

Content-centric Views

From the content perspective, we represent each user with the widely-used bag-of-words model, where each user is represented by a vector in the vector space model. To do so, we extract all the words from twitter messages and construct a user-word matrix, i.e., $\mathbf{V}_6$, where its i-th row $\mathbf{V}_{6(i*)}$ denotes the i-th user's vector which is constructed by Tf-Idf model. As textual features are known to be noisy and the feature space is large, we remove common stop words and only retain words which has been used more than two times to improve the quality of the extracted features, as suggested by several studies (Li et al., 2015, Tang and Liu, 2012, Yang et al., 2014b)

Semantic Views

Tags are known as user-defined keywords demonstrating the underlying topics and semantic concepts exist in the text (Lu, Chen, and Park, 2009). As such, many existing studies have utilized the tags occurred in the text as a semantic feature and description (Lu, Chen, and Park, 2009, Wu et al., 2009). Inspired by these research efforts, we consider hashtags as instances of semantic concepts of the text, which demonstrate the user's interest. We thus construct a user-hashtag matrix,

i.e., $\mathbf{V}_7$ in which each row $\mathbf{V}_{7(i*)}$ represents a vector of hashtags used by the i-th user in status messages.

6.5 Multi-View Profile Learning

In this section, we first introduce an approach to integrate different social behaviours of the users into a unified latent space for user profiling (Section 6.5.1 and Section 6.5.2), and then illustrate how to discover communities (Section 6.5.3) and guide community discovery process with prior knowledge (Section 6.5.4). We finally introduce a pooling approach for constructing community profile from its members (Section 6.5.5).

6.5.1 Preliminaries

Prior studies have shown that social media features can be projected to a latent space with a lower dimensionality, resulting in a dense representation of the original features (Li et al., 2015, Song et al., 2015a). This factorization process is capable of reconstructing the observed entries of original matrix with an effective approximation, i.e., user-item matrix in recommender systems. Inspired by these research findings, we utilize nonnegative matrix factorization (NMF) to decompose users' information into two low rank matrices which are capable of approximately reconstructing the observed matrix. Assume that $\mathbf{V} \in \mathbb{R}^{n \times f}$ represents the data matrix of non-negative elements, the aim of factorization is to decompose $\mathbf{V}$ into two non-negative matrices $\mathbf{W} \in \mathbb{R}^{n \times l}$ and $\mathbf{S} \in \mathbb{R}^{f \times l}$, whose product provide a

good approximation of $\mathbf{V}$, i.e., $\mathbf{V} \approx \mathbf{W}\mathbf{S}^T$. Note that $l \ll f$ is the a pre-specified parameter denoting the dimension of the latent space. Formally, NMF aims at minimizing the following objective function,

$$\min_{\mathbf{W},\mathbf{S}} \|\mathbf{V} - \mathbf{W}\mathbf{S}^T\|_F^2 \quad \text{s.t.} \quad \mathbf{W} \geq 0, \mathbf{S} \geq 0, \tag{6.5}$$

where $\mathbf{W}$ is called the *latent representation of users* and $\mathbf{S}$ is the *latent representation of features* in the low-dimensional latent space. Both $\mathbf{W}$ and $\mathbf{S}$ are non-negative matrices to be learned.

6.5.2 Multi-View Profile Learning

In typical user profiling techniques, users are embedded in a low-dimensional space representing the latent profile of the users, i.e., $\mathbf{W}$. As we have multiple social views corresponding to distinct social behaviors of the user, it is feasible to construct the user profile based on each behavior. More specifically, given m distinct views denoted as $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_m$, each view can be factorized as $\mathbf{V}_i \approx \mathbf{W}_i\mathbf{S}_i^T$, where $\mathbf{W}_i \in \mathbb{R}^{n \times l}$ and $\mathbf{S}_i \in \mathbb{R}^{f_i \times l}$, where f_i denotes the number of features in the i-th social view.

The hypothesis behind multi-view user profiling is that the latent embedding of users from different views should be consistent to each other. Mathematically,

it can formally be stated as the following objective function,

$$J_{user} = \sum_{i=1}^{m} \lambda_i \|\mathbf{V}_i - \mathbf{W}\mathbf{S}_i^T\|_F^2, \quad \text{s.t.} \quad \mathbf{W} \geq 0, \mathbf{S}_i \geq 0, \tag{6.6}$$

where $\mathbf{W}$ is the extracted embedding of users from the fusion of different social views, $\mathbf{S}_i$ is the embedding of features from the i-th social view, m is the number of extracted views, and λ_i is a weight parameter which modulates the effect of the i-th view.

6.5.3 Community Discovery

Recently Yang et. al (Yang et al., 2015) demonstrated that existing community detection algorithms, such as spectral clustering, and modularity, can be interpreted as a clustering of network's nodes in a latent space, where the members of each community form a distinct cluster. Further, similar reserach efforts have demonstrated that each latent dimension in the user profile represents a certain aspect of user interest or behavior. These two observations enlighten us to cluster users in the latent space instead of nodes in the network structure.

Upon the above discussion, we assume that the users form $k \ll n$ communities, and the j-th cluster is defined as $\mathcal{I}_j = \{i \mid i \in \text{community} \quad j\}$. We capture the optimal community structure of users by minimizing the sum of square differences between profiles of users in each community, where differences are computed in the latent space, analogous to major clustering techniques such as K-means and spec-

tral clustering. Let $\bar{\mathbf{W}}_{(j*)} = \frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\mathbf{W}_{(i*)}$ denote the mean of the j-th cluster, and n_j denotes the number of users in the j-th community. Hence, the optimal community structure can be discovered by minimizing the following objective function (Jain, 2010),

$$\min_{\mathcal{I}} J_{com} = \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\|\mathbf{W}_{(i*)} - \bar{\mathbf{W}}_{(j*)}\|_2^2, \tag{6.7}$$

where $\mathbf{W}_{(i*)}$ is the latent representation of the i-th user, and Eq. (6.7) tends to discover user groups with the highest intra-similarities. Similar to (Zha et al., 2001), Eq. (6.7) can be rewritten as follows,

$$J_{com} = tr(\mathbf{W}\mathbf{W}^T) - tr(\mathbf{H}^T\mathbf{W}\mathbf{W}^T\mathbf{H}) \quad \text{s.t.} \quad \mathbf{H}^T\mathbf{H} = \mathbf{I}, \tag{6.8}$$

where $\mathbf{H}$ is the cluster indicator matrix and $\mathbf{H}_{ij} = \frac{1}{\sqrt{n_j}}$ if the i-th user is a member of j-th cluster, otherwise 0. If we ignore the special structure on $\mathbf{H}$ and only keep the orthogonality requirement, the relaxed minimization problem can be mathematically formulated as follows,

$$\min_{\mathbf{W},\mathbf{H}:\mathbf{H}^T\mathbf{H}=I} J_{com} = tr(\mathbf{W}\mathbf{W}^T) - tr(\mathbf{H}^T\mathbf{W}\mathbf{W}^T\mathbf{H}), \tag{6.9}$$

where $\mathbf{H} \in \mathbb{R}^{n\times k}$ demonstrates the affiliation matrix of users so that its ij-th entry, i.e., $\mathbf{H}_{(ij)}$, represents the membership of the i-th user to the j-th community.

6.5.4 Incorporation of Prior Knowledge

It is well-known that the performance of community detection algorithms is limited in situations with excessive noise and missing data (Leskovec, Lang, and Mahoney, 2010). Further, in many real cases, users may have partial prior knowledge or specific community of interest in mind. This prior knowledge can be effectively utilized to boost the performance of community discovery. Although different type of domain knowledge may be available, we focus on pairwise priors since alternative structures (e.g., tree, graph) can simply be converted to pairwise constraints (Wang, Qian, and Davidson, 2012). In previous section, we transformed the community detection into a clustering process in the latent space. In this section, we propose an approach to incorporate domain knowledge into the community discovery process.

Generally, positive and negative pairwise priors are two common type of prior knowledge about community structure. A positive pair (u_i, u_j) attests that the two users should have similar community affiliations, while a negative prior declares different communities. Intuitively, we transfer prior knowledge into the latent space by imposing constraints in affiliation vectors of the users. Specifically, we assume that the affiliation vector of positive pairs are close to each other while the negative pairs are far from each other. Thus these constraints can be mathematically formulated as minimizing the following objective function,

$$J_{prior} = \sum_{(u_i,u_j)\in\mathcal{O}_+} \|\mathbf{H}_{(i*)} - \mathbf{H}_{(j*)}\|_2^2 - \sum_{(u_i,u_j)\in\mathcal{O}_-} \|\mathbf{H}_{(i*)} - \mathbf{H}_{(j*)}\|_2^2$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{Z}_{ij}\|\mathbf{H}_{(i*)} - \mathbf{H}_{(j*)}\|_2^2, \tag{6.10}$$

where $\mathbf{Z} \in \mathbb{R}^{n\times n}$ is the pairwise prior matrix and is precalculated by the following definition,

$$\mathbf{Z}_{ij} = \begin{cases} +1, & (u_i, u_j) \in \Omega_+, \\ -1, & (u_i, u_j) \in \Omega_-, \\ 0 & \text{no prior is available.} \end{cases} \tag{6.11}$$

We can rewrite the Eq. (6.10) as follows,

$$J_{prior} = tr(\mathbf{H}^T(\mathbf{D} - \mathbf{Z})\mathbf{H}) = tr(\mathbf{H}^T\mathcal{L}\mathbf{H}) \tag{6.12}$$

where $\mathcal{L} = \mathbf{D} - \mathbf{Z}$ is the Laplacian matrix (Joachims and others, 2003), and $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with $\mathbf{D}_{(ii)} = \sum_{j=1}^{n} \mathbf{Z}_{(ij)}$.

6.5.5 Community Profiling

The ultimate goal of community profiling is to seek attributes and features which are held by the majority of the group. Thus, we estimate the profile of a community by aggregating of its member profiles, which boils down to the following pooling

operation,

$$\mathbf{G}_{j*} = \sum_{i=1}^{n} \mathbf{W}_{i*}\mathbf{H}_{ij}, \tag{6.13}$$

where $\mathbf{G}_{j*} \in \mathbb{R}^{1\times l}$ is the latent profile of the j-th community. Indeed, by finding the affiliation matrix of users, computing the group profile in the latent space is a straightforward and a linear operation. This is an interesting property which enables us to directly compare users and communities in the latent space, assisting us in solving the "cold-start" problem for new incoming users and those with sparse information.

6.6 Unified Framework

In our unified framework, we fuse the aforementioned collective behavior co-factorization, user clustering in the latent space, and domain knowledge, which can be formalized as the following minimization problem:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{S}_i,\mathbf{H}} J = \sum_{i=1}^{m} \lambda_i \|\mathbf{V}_i - \mathbf{W}\mathbf{S}_i^T\|_F^2 + \alpha\Big(tr(\mathbf{W}\mathbf{W}^T) - tr(\mathbf{H}^T\mathbf{W}\mathbf{W}^T\mathbf{H})\Big) \\ + \beta tr(\mathbf{H}^T\mathcal{L}\mathbf{H}) \quad \text{s.t.} \quad \mathbf{W} \geq 0, \mathbf{S}_i \geq 0, \mathbf{H} \geq 0\end{aligned} \tag{6.14}$$

where α, and β are regularizer parameters to control the tradeoff between different components.

6.6.1 Alternating Optimization

The objective function defined in Eq. (6.14) is not convex with respect to the three variables $\mathbf{W}$, $\mathbf{S}_i$, and $\mathbf{H}$, simultaneously. Hence, there is no closed-form solution for the problem. We adopt alternating optimization approach to find the optimal values for the model parameters. In particular, we alternatively optimize one variable while keeping the others fixed. To enforce the non-negativity constraints, we need to incorporate Lagrangian multipliers. Let Δ_w, Δ_{si}, and Δ_h, be the lagrange matrices for constraints $\mathbf{W} \geq 0$, $\mathbf{S}_i \geq 0$, and $\mathbf{H} \geq 0$, respectively, we rewrite the Lagrange $\mathcal{J}$ as,

$$\begin{aligned}\min_{\mathbf{W},\mathbf{S},\mathbf{H}} \mathcal{J} = \sum_{i=1}^{m} \lambda_i \|\mathbf{V}_i - \mathbf{W}\mathbf{S}_i^T\|_F^2 + \alpha\Big(tr(\mathbf{W}\mathbf{W}^T) - tr(\mathbf{H}^T\mathbf{W}\mathbf{W}^T\mathbf{H})\Big) \\ + \beta tr(\mathbf{H}^T\mathcal{L}\mathbf{H}) + tr(\Delta_w\mathbf{W}) + \sum_{i=1}^{m} tr(\Delta_{si}\mathbf{S}_i) + tr(\Delta_h\mathbf{H}). \qquad (6.15)\end{aligned}$$

We first optimize $\mathbf{W}$ while keeping $\mathbf{S}_i$, and $\mathbf{H}$ fixed. Taking the derivative with respect to $\mathbf{W}$, we have,

$$\frac{\partial \mathcal{J}}{\partial \mathbf{W}} = \sum_{i=1}^{m} 2\lambda_i[\mathbf{W}\mathbf{S}_i^T\mathbf{S}_i - \mathbf{V}_i\mathbf{S}_i] + 2\alpha\mathbf{W} - 2\mathbf{H}\mathbf{H}^T\mathbf{W} + \Delta_w. \qquad (6.16)$$

Using the Karuch-Kuhn-Tucker (KKT) complementary condition, we have the

following update rule for $\mathbf{W}$,

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\sum_{i=1}^{m} \lambda_i \mathbf{V}_i \mathbf{S}_i}{\sum_{i=1}^{m} (\lambda_i \mathbf{W} \mathbf{S}_i^T \mathbf{S}_i) - \mathbf{H}\mathbf{H}^T\mathbf{W} + \alpha\mathbf{W}} \tag{6.17}$$

Similarly, we can obtain the derivation and update rule for $\mathbf{S}_i$ as follows,

$$\frac{\partial \mathcal{J}}{\partial \mathbf{S}_i} = -2\lambda_i \mathbf{V}_i^T \mathbf{W} + 2\lambda_i \mathbf{S}_i \mathbf{W}^T \mathbf{W} + \Delta_{si} \tag{6.18}$$

$$\mathbf{S}_i \leftarrow \mathbf{S}_i \odot \frac{\lambda_i \mathbf{V}_i^T \mathbf{W}}{\lambda_i \mathbf{S}_i \mathbf{W}^T \mathbf{W}} \tag{6.19}$$

Finally, we can rewrite the cost function as follow,

$$\begin{aligned} \mathcal{J} &= \alpha tr(\mathbf{H}^T \mathbf{W} \mathbf{W}^T \mathbf{H}) + \beta tr(\mathbf{H}^T \mathcal{L} \mathbf{H}) + C \\ &= \alpha tr\Big(\mathbf{H}^T (\mathbf{W}\mathbf{W}^T - \xi\mathcal{L})\mathbf{H}\Big) + C \end{aligned} \tag{6.20}$$

where C is constant with respect to $\mathbf{H}$, and $\xi = \frac{\beta}{\alpha}$. With the following *Ky Fan* theorem, we know that the optimal solution of Eq. (6.20) would be $\mathbf{H}^* = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_k]$, where $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_k$ are the eigenvectors of matrix $\mathbf{W}\mathbf{W}^T - \xi\mathcal{L}$.

Theorem 6.1. *(Ky Fan) (Zha et al., 2001). Let* $\mathbf{Q} \in \mathbb{R}^{d \times d}$ *be a symmetric matrix with eigenvalues*

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \tag{6.21}$$

and the corresponding eigenvectors $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d]$. *Then*

$$\lambda_1 + \lambda_2 + \ldots + \lambda_d = \max_{\mathbf{X}^T\mathbf{X}=\mathbf{I}_k} tr(\mathbf{X}^T\mathbf{Q}\mathbf{X}). \tag{6.22}$$

Moreover, the optimal $\mathbf{X}^* = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d]$ *subject to orthonormal transformation.*

6.7 Experiments

6.7.1 Evaluation Metrics

Several metrics have been proposed in previous studies to evaluate the quality of communities. The evaluation metrics can be divided in two categories: quality metrics and consensus metrics. We evaluated the quality of the extracted communities from both aspects.

For quality metrics, we utilized two widely-used metrics named *Davies-Bouldin index*(**dbi**), and *silhouette* (**sil**). The **dbi** metric computes the ratio of the within cluster scatter to the between cluster separation, and hence a lower value means better clustering result. Let c_i be the center of the i-th cluster, and $d(c_i, c_j)$ denotes the distance between centroid c_i and c_j, and σ_i is the average distance between all members of the i-th cluster and c_i. Then **dbi** metric is computed as follows,

$$\mathbf{dbi}(\mathcal{C}) = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right). \tag{6.23}$$

The **sil** score, however, indicates the degree of similarity of a member to its own community compared to other communities, where higher value indicates better clustering and it is defined as,

$$\mathbf{sil}(\mathcal{C}) = \frac{1}{k}\sum_{i=1}^{k}(\frac{1}{|\mathcal{C}_i|}\sum_{i\in\mathcal{C}_i}\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}), \tag{6.24}$$

where

$$a(i) = \frac{1}{|\mathcal{C}_i|-1}\sum_{j\in\mathcal{C}_i, i\neq j} d(i,j), \quad b(i) = \min_{j,j\neq i}\frac{1}{|\mathcal{C}_j|}\sum_{j\in\mathcal{C}_j} d(i,j). \tag{6.25}$$

While quality metrics measure the performance based on the data itself, consensus metrics evaluate the performance of community detection based on external information that was not used in the clustering such as the known class labels and an alternative gold clustering. For consensus metrics, we utilized *normalized mutual information* (**nmi**), and *variation of information* (**vi**). In information theory, the **vi** metric computes the amount of information we obtain when going from one clustering to the other clustering, and it is formally defined as,

$$\mathbf{vi}(\mathcal{C},\mathcal{C}') = H(\mathcal{C}) + H(\mathcal{C}') - 2MI(\mathcal{C},\mathcal{C}'), \tag{6.26}$$

where $H(\mathcal{C})$ denotes the entropy of a clustering $\mathcal{C}$, and $MI(\mathcal{C},\mathcal{C}')$ denotes mutual information between $\mathcal{C}$ and $\mathcal{C}'$. Intuitively, a lower value of **vi** represents better

clustering. Similarly, **nmi** is computed as,

$$\mathbf{nmi}(\mathcal{C}, \mathcal{C}') = \frac{MI(\mathcal{C}, \mathcal{C}')}{\sqrt{H(\mathcal{C})H(\mathcal{C}')}}, \tag{6.27}$$

where a higher value indicates better clustering.

6.7.2 On Model Performance Comparison

To demonstrate the effectiveness of our approach, we compared our community detection method with the following state-of-the-art methods,

K-Means: **K-Means** is a traditional clustering method which is widely used for discovering communities in social networks based on users' social connections (Qi, Aggarwal, and Huang, 2012).

N-Cut: **N-Cut** is the clustering method based on normalized cut proposed in (Shi and Malik, 2000) which detects communities based on the network structure between users.

Pair-Spec: **Pair-Spec** extends the well-known spectral clustering method into multi-view setting by co-regularizing the clustering hypothesis across views (Kumar, Rai, and Daume, 2011).

SI: Zhou et. al (Zhou and Liu, 2013, Zhang and Yu, 2015) recently proposed a method for community detection based on social influence. **SI** propagates heterogenous information across views to calculate the co-influence of users

Table 6.3: Community detection results for different approaches in terms of quality metrics (first two rows) and consensus metrics(last two rows).

Methods	**dbi**	**sil**	**vi**	**nmi**
K-Means	0.912	-0.340	5.941	0.120
N-Cut	0.860	-0.387	6.221	0.115
Pair-Spec	0.855	-0.342	4.341	0.146
SI	0.824	-0.219	5.347	0.123
MCD	0.790	-0.118	4.191	0.147
Latent	0.782	-0.022	2.875	0.175

based on the social graph and its associated activity graphs.

MCD: Mutual clustering discovery (**MCD**) is proposed to detect communities of users across multiple networks based on a meta-path similarity measure. We changed the method proposed in (Yu and Zhang, 2015) to compute the communities from multiple views of the social network.

Latent: Latent is our proposed community discovery approach, i.e., Eq. (6.14), where communities are discovered in the latent space. As these baseline cannot utilize prior knowledge, to have a fair evaluation, we did not use prior information in this section (set $\beta = 0$).

Table 6.3 shows the clustering results of different methods in terms of the introduced metrics in Section 6.7.1. From the table, the following points can be observed. In terms of community quality: (1) **Latent** achieves the lowest **dbi** and the highest **sil** amongst all the community discovery approaches, demonstrating that our model is able to detect highly dense clusters in the latent space with proper discrimination between distinct communities. (2) **K-Means** and **N-Cut** achieve neck to neck performance with a slight improvement by **N-Cut**, which is

consistent with previous studies (Yu and Zhang, 2015). This is due to the fact that **N-Cut** is intrinsically defined to partition the graph into dense subgraphs. (3) **Pair-Spec** obtains better result than **N-Cut** verifying the importance of other information sources in the community discovery process. Similar findings have also been reported in retrospective research efforts (Kumar, Rai, and Daume, 2011). **SI** and **MCD** outperform their counterparts with **MCD** has better performance in both metrics, which is attributed to the information fusion techniques in the two approaches. **SI** analyzes each of the views separately and transforms information from one to the other, which may fail to capture the correlation between the views. **MCD**, on the other hand, models the heterogenous information in multiple views using meta-paths by considering all possible correlations between the views, as reported by (Yu and Zhang, 2015).

Similar results can be observed from consensus metrics under the evaluation of **vi** and **nmi** metrics. For example, **K-Means**, and **N-Cut** have the lowest **nmi**, emphasizing that the network-information fails to find the underlying community structure on social networks due to its sparsity (Leskovec, Lang, and Mahoney, 2010, Yang et al., 2009). Further, **Pair-Spec**, **SI**, and **MCD** achieve better performance by utilizing other side information. Last but not least, **Latent** model achieves lower **vi** and higher **nmi**, which shows the ability of the proposed model in discovering dissimilar communities.

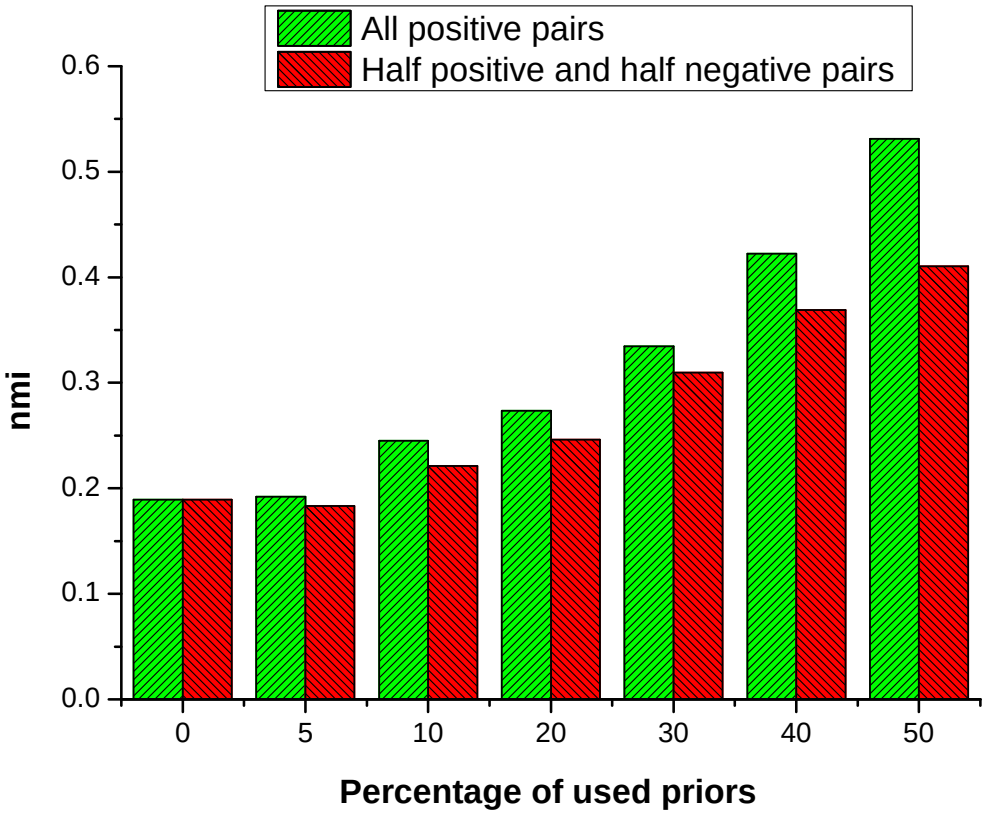


Figure 6.2: The Effect of incorporation of prior knowledge in community extraction, which clearly indicate a positive correlation between amount of prior knowledge and the performance of community discovery. Further, positive constraints contribute more in performance rather than negative priors.

6.7.3 On Incorporation of Prior Knowledge

In this section, we conducted experiments to evaluate the performance of the framework in utilizing prior information in the community discovery process. As the amount of prior information is an important factor in semi-supervised learning (Ver Steeg, Galstyan, and Allahverdyan, 2011, Zhang, Sun, and Wang, 2013), we randomly sampled a set of users to construct the positive and negative pairs of prior information. To construct prior constraints, we randomly selected two users. If these two users belong to the same community, they form a positive pair, otherwise a negative pair. The community label of each user was inferred based on the ground-truth labels extracted from the social network (Section 6.4). The number of communities was set to the ground-truth's community number, i.e. $k = 10$, where other parameters were set to their best value based on the parameter study (See Section 6.7.4).

Figure 6.2 demonstrates the performance of our method corresponding to different percentages of prior information used. As can be seen from the Figure, there exists a clear positive correlation between the value of **nmi** and the amount of prior pairs used, where the effect of positive pairs is clearly stronger than the effect of negative prior pairs. This is owning to the fact that positive priors precisely describe the local community structure as compared to negative priors which avoid misclassification. Further, the effect of prior information is not obvious when the amount of prior knowledge is low, i.e. less than 10%. Overall, guiding the commu-

nity discovery process with prior information results in better performance, where the positive priors contribute more in the final performance. Similar results have been observed for the **vi** metric.

6.7.4 On Parameter Tuning

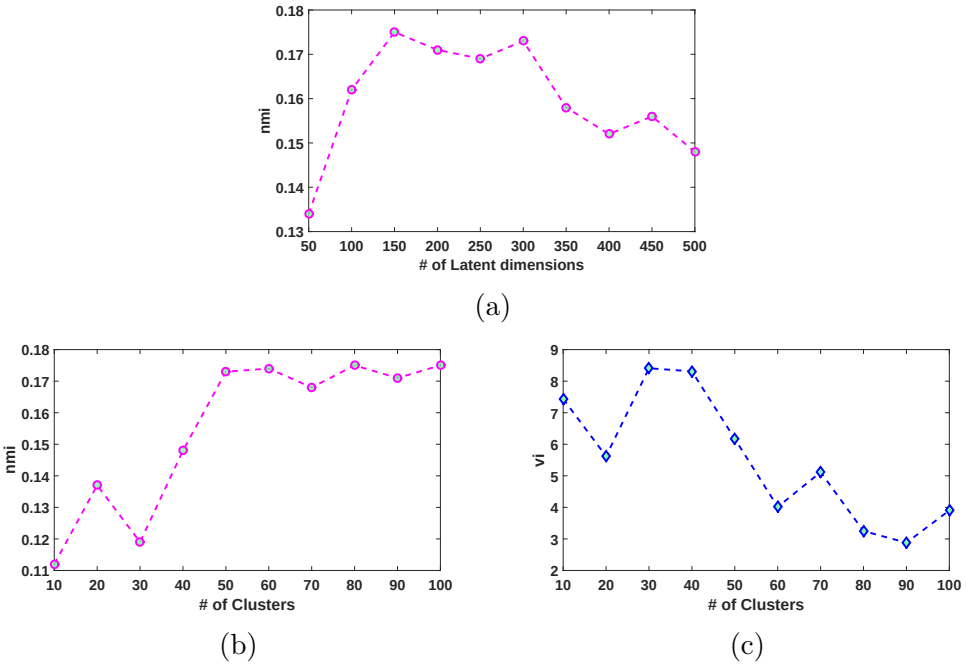


Figure 6.3: Effect of model's hyper-parameters. (a) shows the effect of the latent dimension, l. (b) and (c) show the effect of the number of clusters, k. The stability of performance for values above 50 demonstrates that there exists several sub-communities in each social support group.

We also studied the parameter sensitivity of our proposed framework. Our framework holds two sets of parameters: (1) the model hyper-parameters which define two distinct aspects of the model: the latent space dimension l and the number of clusters k; (2) the regularizers α, β, and λ_i in Eq. (6.14), which control the effect of different components. We first evaluated the sensitivity of the proposed approach to the dimension of the latent space and the number of clusters

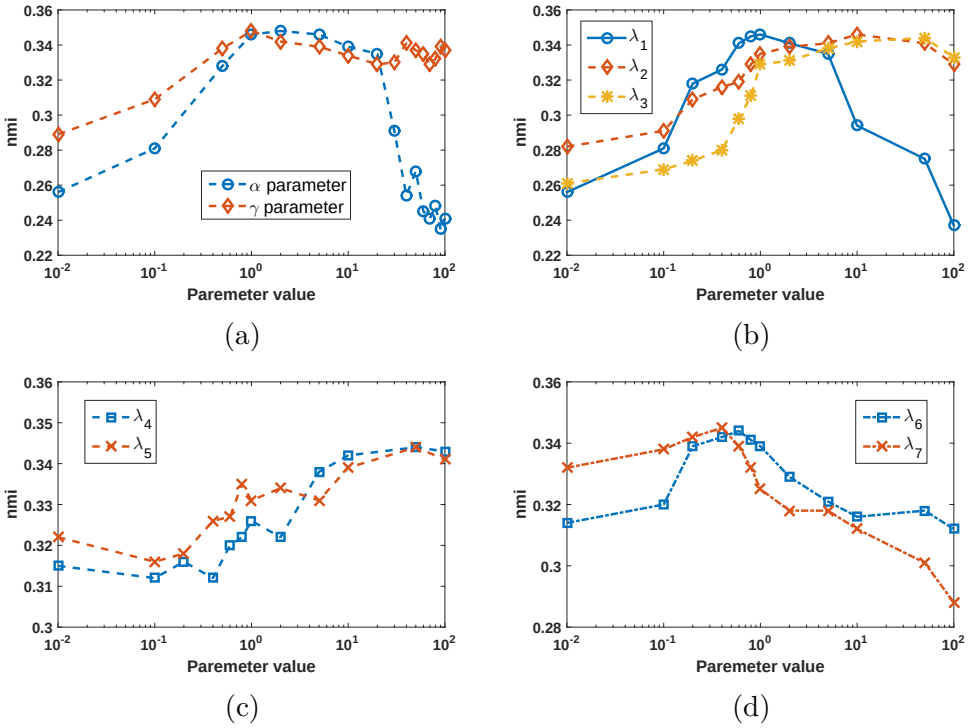


Figure 6.4: Effects of different regularizer parameters in community discovery. (a) The effect of regularizer weight for different components of the model. (b) to (d) The effect of weight value for different social views.

to see how these parameters affect the community discovery process. We next evaluated the performance of the proposed approach with respect to regularizers.

To assess the effect of the dimension of the latent space, we varied l in the range of $\{50, 100, \ldots, 500\}$, while keeping the other parameters fixed. Figure 6.3a illustrates the performance of the community discovery process in terms of **nmi**. The performance reaches the best value when the number of latent dimensions is around 150. The results show that when the number of latent dimension is too small, the model is unable to find effective communities. In contrast, a large latent dimension will result into the loss of performance, probably due to overfitting phenomenon. This result is also consistent with a large body of research in feature

learning which attests the importance of proper feature learning (Li et al., 2015, Tang and Liu, 2012).

To study the effect of the parameter k which determines the number of clusters, we conducted the experiment with various values for $k = \{10, 20, \ldots, 100\}$ while keeping the latent dimension fixed to its best value, i.e., $l = 150$. Figure 6.3b, and 6.3c show the performance of the proposed framework in terms of both **nmi** and **vi**, respectively. As can be seen from the Figure, the results achieved by our method is very stable for k values above 50 for both **nmi**, and **vi**. It is worth noting that the instability of **nmi** for values below 50 shows the targeted support communities can be divided into several sub communities with focused interests.

We next studied the effect of α and β which are regularizer parameters to modulate the effect of different components, as shown in Figure 6.4a. From the Figure, we observe that low values of α and β cannot find effective clusters in the latent space as well as fail to incorporate prior knowledge into the community discovery process, respectively. The reason is that with low values of these parameters the model tends to scatter the users in the latent space to learn an effective embedding for each users, failing to find similarities amongst users to construct communities. In other words, the learning process focuses to minimize the first term in Eq. (6.14) and overlooks the remaining components. Similarly, large values of α have a negative effect making all the embedding of the users are located near to each other to have dense clusters, failing to find existing contrasts between different clusters.

There is another set of parameters $\{\lambda_i\}$ in the proposed method, which controls the relative importance of each view in computing the latent space. We evaluated the effect of each parameters in the final performance of the framework. To do so, we selected one parameter at each time and varied its values in the range of $\{0.001, 0.01, \ldots, 100\}$ while setting the others to their best values. The results are shown in Figure 6.4b to 6.4d, where we categorized them based on their effects on the final performance. From the Figure, we can observe that the performance of the model is stable with respect of the parameters, while the optimal value is different for each parameter. The best value for the network information, i.e., $\mathbf{V}_1$, which traditionally is the main information source for community detection is, $\lambda_1 = 1$, when more sparse views like $\mathbf{V}_2$, and $\mathbf{V}_3$ hold a larger weight, i.e., $\lambda_2, \lambda_3 \in [10 - 50]$. Meanwhile, noisy views with a lot of redundant features obtain a lower weight $\lambda_6 = 0.2$, and $\lambda_7 = 0.5$, probably to perform feature selection. As far as we know, while multi-source and multi-view research in social network has recently attracted much research, there is a limited research on quantifying the importance of different social views. This finding itself demonstrates the importance of distinct behavior of the user as well as provides insights about their relative importance in community discovery. Further research in this direction is important and can reveal invaluable insights.

6.7.5 Qualitative Study

While macro-level quantitative evaluation is useful, it is also instructive to examine actual results to better understand the outputs of the **Latent** model in community profiling. To accomplish this end, we first investigate the interpretability of extracted latent features from the data, and next give some case studies drawn from the dataset to demonstrate its ability in extracting profilable communities.

Interpretation of the Latent Features

In the **Latent** model, $\mathbf{W}$ is the low-dimensional latent space, while each $\mathbf{S}_i$ serves as the basis matrix for representing a view. Thus, we are able to explain the interpretation of the learnt embeddings from the data by mapping them into the original features. To accomplish this end, we first normalized the weights of the rows in $\mathbf{S}_i$ such that the sum of each row was equal to 1. We then ranked the latent features according to their normalized weights and found the representative latent dimension corresponding to each social views (He et al., 2014b). In Table 6.4 we showed lists of words that were mapped to the leading basis in the content space $\mathbf{S}_6$. For readability, we manually labelled each row. The result verifies that the latent dimensions are constructed by grouping meaningful and correlated words from the published posts. For example, the first dimension represents a word group related to negative emotion and feeling which is typical in people struggling with their health condition (Eysenbach et al., 2004). Similarly, many symptoms and side-effects of medication are clustered into a separate dimension (the last

row), representing a topic pertaining to consumer health information. The results are consistent with several studies in social sciences and health informatics, which reveal that patients are coming online to fulfill their health demands, where information and emotional support are the main themes (Eysenbach et al., 2004). Interestingly, our analysis on **BG** dataset in chapter 4 also demonstrated similar results (See Section 4.5).

Case Study

To have a tangible understanding of the outcome of our method, we show examples of discovered communities with their profiles in Table 6.5. For each community, we show top latent features in terms of prominent words, hashtags, and leading users. As shown in the first row, spammers frequently publish advertisements related to devices (#insulin pump, and glucometer) or nutritional supplements (herbal, supplement, and #weightloss). Further, a closer inspection of twitter accounts u_1794, and u_5237[4] reveals that these two accounts frequently mention products related to diabetes. Similarly, health seekers also use twitter to acquire real-time information about their health questions (the second row), such as "Is there any cure for diabetes?", and "Why do my legs hurt when I start walking?", while the former asked by a user who have been active for few days, probably recently diagnosed, and the second question published by "@diabetes_sanofi" along with a web page address. Indeed, $@diabetesdaily$, $@diabetes\_sanofi$ are diabetes advocates groups which often publish authorized health information in Twitter.

[4]Due to the privacy issues, we have anonymized these users.

Table 6.4: Sample leading latent features drawn from the content social view.

ID	Name	Top 5 Words
43	Negative emotion	tired, horrible, hurt, missed, shit
72	Medications	insulin, pump, actos, avanda, metformin
116	Social support	ask, help, advice, sweet, easy
127	Activities	running, gym, daily, work, runtastic
130	Symptoms	hypertension, thirst, neuropathy, disease, cataract

Table 6.5: Sample community profiles in terms of prominent words, hashtags, and leading users.

Name	Top words	Top hashtags	Leading users
Spammers	weight, herbs, less, supplement, review, beat, Glucometer	#weightloss, #insulin	u1794, u1001, u5237
Health seekers	care, treatment, advice, change, meal, hyperglycaemia, foot, depression	#diabetestips, #doc, #dsma	@askUHC, @diabetesdaily, @diabetes_sanofi

Similarly, $@askUHC$ is also an important social feature for this community since it is the twitter account of UnitedHealth Group, which answers patients question online, and it was introduced by one member several times during the crawling period.

6.8 Related Work

Discovering communities has been of interest by many retrospective studies in complex network analysis (Newman and Girvan, 2004). Existing approaches can be categorized into two groups: measure-based approaches and probabilistic approaches. Typically, measure-based approaches define an objective function to quantify the quality of a cluster with the assumption that a good cluster is a set

of nodes with dense internal connectivity and sparse external connectivity (Newman and Girvan, 2004). Upon this intuition, several quality measure have been proposed in the literature to evaluate the clusters based on link and content information such as normalized cut (Shi and Malik, 2000) and modularity (Newman and Girvan, 2004). Probabilistic approaches, such as stochastic block model (Abbe and Sandon, 2015) and its variants (Hofman and Wiggins, 2008), assume that links are generated with the probability that only depends on the communities of the nodes. However, due to the excessive noise in content and sparsity in link information, neither link information, nor content information is sufficient for inferring the optimal community structure (Yang et al., 2009, Zhuang et al., 2015). Thus, combining link and content information has been utilized to improve the performance of community detection such as Link-LDA (Erosheva, Fienberg, and Lafferty, 2004), kernel fusion (Yu, De Moor, and Moreau, 2009), and PHITS-PLSA (Cohn and Hofmann, 2001). However, most of these approaches apply a generative model for content analysis which is sensitive to noisy information in social networks. Moreover, most of these approaches fail to provide an interpretable community profile.

In the area of semi-supervised learning, several community detection approaches have been proposed recently (Eaton and Mansbach, 2012,Ma et al., 2010,Ver Steeg, Galstyan, and Allahverdyan, 2011). For example, spin-glass model from statistical physics has been utilized to incorporate prior knowledge into community detection (Eaton and Mansbach, 2012). Similarly, several research efforts proposed to

directly change the adjacency matrix in order to apply prior knowledge in community detection process (Ma et al., 2010, Zhang, 2013, Zhang, Sun, and Wang, 2013). However, directly modifying the adjacency matrix cannot guarantee that the two nodes belong to the same community as reported by (Ma et al., 2010, Zhang, Sun, and Wang, 2013). In contrast to prior investigations, our aim is to fuse all the social views available in the network for discovering user communities as well as provide a latent profile for each extracted communities, which can be explained in the feature space.

6.9 Summary

In this chapter, we proposed a community discovery and profiling approach for social media users. The proposed model simultaneously learns the profiles of users and their affiliations to communities in a low-dimensional space, which is constructed from the integration of different social views of the network. In particular, we integrated different social views of the network into a low-dimensional latent space in which we sought dense clusters of users as communities. By imposing a Laplacian regularizer into affiliation matrix, we further incorporated prior knowledge into the community discovery process. Finally community profiles were computed by a linear operator integrating the profiles of members.

This study have demonstrated: (1) the importance of learning the community structure from all social behaviors of users, which can be achieved by learning a unified latent space; (2) the significance of incorporating prior knowledge in

community discovery and profiling; and (3) an approach to learn the profile of communities, which helps understand the collective behavior of users.

CHAPTER 7

Conclusion and Future Work

7.1 Conclusion

In this thesis, we aimed to make sense of wellness of users on social networks, both at micro-level of individuals, i.e., user profiling, and macro-level of groups and communities, i.e., community profiling. In particular, we focused on learning the wellness profile of users, where we exploited their social media information to identify, understand and estimate the wellness attributes and states of users and communities. To accomplish this, we first harvested social media posts of users to extract personal wellness events which directly expose wellness information and attributes of users. We then studied that how online behavior of users can reflect their wellness information and attributes. Specifically, we attempted to differentiate between adapted and non-adapted users based on their social media data. Following this line of research, we proposed to learn the wellness profile of users and communities on social media platform. The investigated approaches, in

this thesis, permit us to better understand the wellness of users and communities, get actionable insight about the wellness of users and communities from social media, and provide better social and information services.

As a first step towards understanding the wellness of user from social media, we proposed a framework for extracting the mentions of personal wellness events from posts of users in Twitter microblogging service. The proposed framework leveraged the content information of microblogging text as well as the relations amongst wellness event categories to categorize events into a wellness taxonomy. This approach permits us to learn both task-specific and task shared features, which significantly boosts the performance of the learning framework. By imposing a sparse constraint on the learning model, we also tackled the problems arising from noise and variation in microblogging texts. Experimental results on a real-world dataset from Twitter have demonstrated the superior performance of our framework in extracting personal wellness events.

We next studied the behavioral distinction between diabetes patients aiming at characterizing two groups of patients: patients who can successfully manage their diabetes, in terms of blood glucose value, and those who fail to manage their diabetes, referred by adapted and non-adapted patient cohorts, respectively. We studied online behaviors of users in terms of linguistic, textual and visual attributes and contents in their online posts. We have observed several characteristics such as negative affective, seeking and sharing supportive contents, and difference in shared visual concepts, which differs adapted and non-adapted users. We discussed

the implication of our finding from clinical aspect and elaborated on the various limitations and ethical issues of using social media in the wellness domain.

To learn the latent profile of users, we proposed an approach which directly learns the embedding from longitudinal data of users, instead of vector-based representation. In particular, we simultaneously learned a low-dimensional latent space as well as the temporal evolution of users in the wellness space. The proposed method takes into account two types of wellness prior knowledge: (1) temporal progression of wellness attributes; and (2) heterogeneity of wellness attributes in the patient population. Our approach scales well to large datasets using parallel stochastic gradient descent. We conducted extensive experiments to evaluate our framework at tackling three major tasks in wellness domain: attribute prediction, success prediction and community detection. Experimental results on two real-world datasets demonstrated the ability of our approach to learn effective user representations.

To learn the profile of user groups, we proposed to discover communities in a low-dimensional latent space in which we simultaneously learned the representation of users and communities. Specifically, our approach leveraged different online behaviors of users as multiple social views, and fused them into a latent space representing all users' behaviors. Community discovery was then performed in the latent space by considering all social behaviors of the users. By guiding the community discovery process with available prior knowledge, we not only was able to discover communities based on the entire user behaviors but also was able

to compute the community profile which helps to explain the collective behavior of the members of the community. Taking the wellness domain as an example, we have conducted experiments on a large scale real world dataset of users publishing tweets about diabetes and its related concepts, demonstrating the ability of our approach in discovering and profiling user communities.

7.2 Future Work

This research begins a new research direction towards connecting social media and health informatics with many downstream applications. In the previous chapters, we have discussed how to harvest social media information towards making sense of the wellness of users and communities. Following the proposed framework, we can envision several directions for future work. We summarize them into two aspects of: mining information and profile learning.

7.2.1 Mining Information

We demonstrated the possibility of mining social media data to extract rich information about the wellness of users. However, social media platform, per se, is a high-velocity streaming information source. Social media data is generated dynamically, where, everyday, users generate new data and features at a rapid pace. For example, in Twitter, more than 320 tweets are produces daily with a large number of slang words and new hashtags. These terms grab the attention of users and become popular and trending in a short time. Therefore, streaming feature

selection is more practical and desirable to rapidly adapt to the changes (Li et al., 2015, Guo et al., 2014).

Wellness event extraction is an initial step to understand the wellness of users from social media data. However, events are accompanied with several attributes providing specific information about the various aspects of the events. Taking the tweet "195.0 #BGNow @ 08:20AM after bike ride 90 minutes" as an example, the tweet refers to a biking event and an examination of blood glucose. While knowing the event type is important and useful, extracting the details of the event provides precise information about the event which is useful for further analysis. We utilized a bootstrapping rule-based approach for extracting events. Sequence labelling with Conditional Random Field (CRF) is another approach which have demonstrated promising results in information extraction studies. Our proposal for future work in this direction is to have a comparative study between these two algorithms and assess their efficiency and effectiveness on extracting attributes of the wellness events.

7.2.2 Profile Learning

In individual user profiling, this study demonstrated the importance of feature learning approaches which are intrinsically designed for longitudinal data. Different extensions of this work can be investigated in future. The first is to utilize the social context around users in a collaborative learning approach. As social media users are linked to each other, incorporation of network-centric information

is a promising direction. Indeed, recent studies attest that using social context of users can improve the performance of prediction in different tasks such as sentiment analysis (Hu et al., 2013), user interest profiling (Yang et al., 2011), recommendation (Geng et al., 2015), trust prediction (Tang et al., 2013), and so on. One can examine the correlation between wellness attributes and states of users in their ego-network to gain better insight about the wellness of users and develop an effective learning framework.

With the proliferation of social networking services, users simultaneously participate in multiple social networks to enjoy their diverse services. For example, more than half of US adults (52%) and a majority of US teenagers (71%) in 2014 use two or more social media services. Indeed, they publish different information, disclose different attributes, and share about different events in each social network. Thus the integration of users' data from multiple social networks would be a promising research direction.

From group profiling aspect, community profiling can be used to investigate the evolution of communities in a dynamic network environment, where communities can grow, merge, and dissolve. Another promising direction is to complement the individual profiles based on their community affiliation and study how it can facilitate recommendation. Identifying the users who play crucial role in group formation and activities is another interesting promising direction, as this study has already presented such capability (See Section 6.7.5)

7.3 Ethics and Limitations

While thinking about designing intervention programs on social network and in general health, it is important to bear in mind that wellness and health data can be extremely sensitive and need to be verified before providing to the user. Finding the authorized and reliable wellness information is a challenging task especially in a noisy platform such as social media with a lot of user generated contents and spams contents. The truthworthy of the information needs to be considered with a proper automatic or semi-automatic way. This can be done by a human in loop procedure to verify the potential risky unreliable information. Further, the design consideration in social platforms should honor the privacy of the affected individuals and abide the proper ethical guidelines ensuring that the intended profit of the intervention exceeds the potential difficulties and risks. To sum, we hope this research open a new avenue to not only detect and help diabetic patients through social platforms, but also understand the collaborative behaviour of different diabetics communities towards designing better healthcare interventions and treatments.

It is worthwhile noting that this study does not make any claim to attributing the social network as an individual platform through which we can obtain a complete understanding of wellness condition of diabetic users and provide a full intervention program. We however attest that patient generated wellness and lifestyle data on social media can be utilized as a complementary source through

which we can sense users' wellness and lifestyle. Social sensors can be utilized as a complementary source of information in combination of the popular concept of quantified-self measuring users' attributes with wearable devices. We caution against using this method as standalone technique for diagnosis and prediction of diabetes. We also note that social media is a noisy and sparse platform where many users may not utilize it for health information explicitly. However, the implicit signals and clues in their social account can provide useful information when aggregated in scale. Moreover, selection bias and confounding factors are always important problems in social and health studies. There are always a cohort of users who are not active users of social media and social sensors would fail to provide a complete perspective about these users. For instance, while social media is used by different age groups in society, majority of social media users are youngsters, which introduces an inherent source of bias in social media studies.

Finally, our findings reveals the richness of patient generated wellness data on social media demonstrating that it can be used in combination of other information sources to obtain a comprehensive understanding of diabetes patient. It also raises several difficult questions for researchers, as mentioned below. How much social media information is reliable in health domain? How precise user's online behaviours reveal his offline attributes and behaviours? And how effective would the designed intervention be, in terms of changing the user behaviour?

Papers arising from this thesis

- Mohammad Akbari, Kavita Venkataraman, Tat-Seng Chua. #BgNow: A Characterization Study of Diabetes Success on Twitter. *To be submit* Journal of the American Medical Informatics Association (JAMIA).
- Mohammad Akbari, Xia Hu, Tat-Seng Chua. Wellness Representation of Users in Social Media: Towards Joint Modelling of Heterogeneity and Temporality. *Submitted to* IEEE Transactions on Knowledge and Data Engineering (TKDE).
- Mohammad Akbari, Tat-Seng Chua. Leveraging Behavioral Factorization and Prior Knowledge for Community Discovery and Profiling. ACM International Conference on Web Search and Data Mining (WSDM), 2017.
- Mohammad Akbari, Xia Hu, Liqiang Nie, Tat-Seng Chua. On the Organization and Retrieval of Health QA Records for Community-based Health Services, International Joint Conference on Artificial Intelligence (IJCAI), BOOM Workshop, Best Paper Award.
- Mohammad Akbari, Xia Hu, Liqiang Nie, Tat-Seng Chua. From Tweets to Wellness: Wellness Event Detection from Twitter Streams. In Proceeding of AAAI Conference on Artificial Intelligence (AAAI), 2016.
- Mohammad Akbari, Liqiang Nie, and Tat-Seng Chua. aMM: Towards adaptive ranking of multi-modal documents. International Journal of Multimedia Information Retrieval (IJMIR), 2015.

References

Abbar, Sofiane, Yelena Mejova, and Ingmar Weber. 2015. You tweet what you eat: Studying food consumption through twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3197–3206. ACM.

Abbe, Emmanuel and Colin Sandon. 2015. Recovering communities in the general stochastic block model without knowing the parameters. In *NIPS*.

Akbari, Mohammad, Xia Hu, Liqiang Nie, and Tat-Send Chua. 2016. From tweets to wellness: Wellness event detection from twitter streams. In *AAAI*.

Akbari, Mohammad, Liqiang Nie, and Tat-Seng Chua. 2015. amm: Towards adaptive ranking of multi-modal documents. *International Journal of Multimedia Information Retrieval*, 4(4):233–245.

Al-Kamha, Reema and David W Embley. 2004. Grouping search-engine returned citations for person-name queries. In *Proceedings of the 6th annual ACM international workshop on Web information and data management*, pages 96–103. ACM.

Alvarez-Conrad, Jennifer, Lori A Zoellner, and Edna B Foa. 2001. Linguistic predictors of trauma pathology and physical health. *Applied Cognitive Psychology*.

Anderson, Ryan J, Kenneth E Freedland, Ray E Clouse, and Patrick J Lustman. 2001. The prevalence of comorbid depression in adults with diabetes a meta-analysis. *Diabetes care*.

Aronson, Alan R. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *AMIA Symposium*.

Association, American Diabetes et al. 2014. Standards of medical care in diabetes. *Diabetes care*.

Attai, Deanna J, Michael S Cowher, Mohammed Al-Hamadani, Jody M Schoger, Alicia C Staley, and Jeffrey Landercasper. 2015. Twitter social media is an effective tool for breast cancer patient education and support: patient-reported outcomes by survey. *JMIR*.

Backstrom, Lars, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW*, pages 61–70. ACM.

Becker, Hila, Mor Naaman, and Luis Gravano. 2011. Beyond trending topics: Real-world event identification on twitter. In *ICWSM*.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Bo, Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social

media data by finding location indicative words. *Proceedings of COLING 2012: Technical Papers*, pages 1045–1062.

Burger, John D, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *EMNLP*, pages 1301–1309. Association for Computational Linguistics.

Burger, John D and John C Henderson. 2006. An exploration of observable features related to blogger age. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 15–20.

Cai, Deng, Xiaofei He, Jiawei Han, and Thomas S Huang. 2011. Graph regularized nonnegative matrix factorization for data representation. *PAMI*.

Cao, Yonggang, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J Cimino, John Ely, and Hong Yu. 2011. Askhermes: An online question answering system for complex clinical questions. *Journal of biomedical informatics*.

Carlson, Andrew, Scott Gaffney, and Flavian Vasile. 2009. Learning a named entity tagger from gazetteers with the partial perceptron. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, pages 7–13.

Cavalin, Paulo R, Luis G Moyano, and Pedro P Miranda. 2015. A multiple classifier system for classifying life events on social media. In *2015 IEEE International Conference on Data Mining Workshop*, pages 1332–1335. IEEE.

Chang, Chia-Hui, Mohammed Kayed, Moheb Ramzy Girgis, and Khaled F Shaala. 2006. A survey of web information extraction systems. *Knowledge and Data Engineering, IEEE Transactions on*.

Chapman, Daniel P, Geraldine S Perry, and Tara W Strine. 2005. The vital link between chronic disease and depressive disorders. *Prev Chronic Dis*.

Che, Zhengping, Sanjay Purushotham, Robinder Khemani, and Yan Liu. 2015. Distilling knowledge from deep networks with applications to healthcare domain. *arXiv preprint arXiv:1512.03542*.

Chen, Chao, Dongsheng Li, Yingying Zhao, Qin Lv, and Li Shang. 2015. Wemarec: Accurate and scalable recommendation through weighted and ensemble matrix approximation. In *SIGIR*.

Chen, Xi, Weike Pan, James T Kwok, and Jaime G Carbonell. 2009. Accelerated gradient method for multi-task sparse learning problem. In *ICDM*.

Chen, Yan, Jichang Zhao, Xia Hu, Xiaoming Zhang, Zhoujun Li, and Tat-Seng Chua. 2013. From interest to function: Location estimation in social media. In *AAAI*.

Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *CIKM*, pages 759–768. ACM.

Cherry, Colin, Hongyu Guo, and Chengbi Dai. 2015. Nrc: Infused phrase vectors for named entity recognition in twitter. *ACL-IJCNLP 2015*.

Chhieng, D, T Day, G Gordon, and J Hicks. 2006. Use of natural language programming to extract medication from unstructured electronic medical records. In *JAMIA*.

Choudhury, Smitashree and Harith Alani. 2014. Personal life event detection from social media.

Chunara, Rumi, Jason R Andrews, and John S Brownstein. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American journal of tropical medicine and hygiene*, 86(1):39–45.

Cohn, David and Thomas Hofmann. 2001. The missing link-a probabilistic model of document content and hypertext connectivity. *NIPS*.

Cruz, Juan David, Cécile Bothorel, and François Poulet. 2013. Community detection and visualization in social networks: integrating structural and semantic information. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):11.

Dalvi, Nilesh, Ravi Kumar, and Bo Pang. 2012. Object matching in tweets with spatial models. In *WSDM*, pages 43–52. ACM.

Davis, Matthew A, Denise L Anthony, and Scott D Pauls. 2015. Seeking and receiving social support on facebook for surgery. *Social Science & Medicine*.

De Choudhury, Munmun. 2015. Anorexia on tumblr: A characterization study. In *Digital Health*.

De Choudhury, Munmun, Scott Counts, and Eric Horvitz. 2013a. Major life changes and behavioral markers in social media: case of childbirth. In *CSCW*, pages 1431–1442. ACM.

De Choudhury, Munmun, Scott Counts, and Eric Horvitz. 2013b. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3267–3276. ACM.

De Choudhury, Munmun, Scott Counts, and Eric Horvitz. 2013c. Social media as a measurement tool of depression in populations. In *ACM Web Science Conference*.

De Choudhury, Munmun, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *CSCW*, pages 626–638. ACM.

De Choudhury, Munmun, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *ICWSM*.

De Choudhury, Munmun, Emre Kiciman, Mark Dredze, Glen Coppersmith, and

Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media.

De Choudhury, Munmun, Meredith Ringel Morris, and Ryen W White. 2014. Seeking and sharing health information online: comparing search engines and social media. In *SIGCHI*.

De Choudhury, Munmun, Sanket S Sharma, and Emre Kıcıman. 2016. Characterizing dietary choices, nutrition, and language in food deserts via social media.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Diao, Qiming, Jing Jiang, Feida Zhu, and Ee-Peng Lim. 2012. Finding bursty topics from microblogs. In *ACL*.

Dickinson, Thomas, Miriam Fernandez, Lisa A Thomas, Paul Mulholland, Pam Briggs, and Harith Alani. 2015. Identifying prominent life events on twitter. In *Proceedings of the 8th International Conference on Knowledge Capture*, page 4. ACM.

Ding, Chris, Tao Li, and Wei Peng. 2008. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8):3913–3927.

Doan, Son, Nigel Collier, Hua Xu, Pham H Duy, and Tu M Phuong. 2012. Recog-

nition of medication information from discharge summaries using ensembles of classifiers. *BMC medical informatics and decision making*.

Dos Reis, Virgile Landeiro and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health from twitter. In *AAAI*.

Dredze, Mark. 2012. How social media will change public health. *Intelligent Systems, IEEE*.

Eaton, Eric and Rachael Mansbach. 2012. A spin-glass model for semi-supervised community detection. In *AAAI*.

Eichstaedt, Johannes C, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, et al. 2015. Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169.

Erosheva, Elena, Stephen Fienberg, and John Lafferty. 2004. Mixed-membership models of scientific publications. *PNAS*.

Eysenbach, Gunther. 2006. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In *AMIA Annual Symposium Proceedings*, volume 2006, page 244. American Medical Informatics Association.

Eysenbach, Gunther, John Powell, Marina Englesakis, Carlos Rizo, and Anita

Stern. 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *BMJ*.

Farseev, Aleksandr, Liqiang Nie, Mohammad Akbari, and Tat-Seng Chua. 2015. Harvesting multiple sources for user profile learning: a big data study. In *ICMR*.

Farseev, Aleksandr, Nickolay Zhukov, Ilia Gossoudarev, and Uri Zarichnyak. 2014. Cross-platform venue recommendation based upon user community detection from social media.

Fogg, BJ and Jason Hreha. 2010. Behavior wizard: a method for matching target behaviors with solutions. In *Persuasive technology*.

Franciosi, Monica, Fabio Pellegrini, Giorgia De Berardis, Maurizio Belfiglio, Donatella Cavaliere, Barbara Di Nardo, Sheldon Greenfield, Sherrie H Kaplan, Michele Sacco, Gianni Tognoni, et al. 2001. The impact of blood glucose self-monitoring on metabolic control and quality of life in type 2 diabetic patients an urgent need for better educational strategies. *Diabetes care*.

Garrett, Daniel G and Benjamin M Bluml. 2005. Patient self-management program for diabetes: first-year clinical, humanistic, and economic outcomes. *Journal of the American Pharmacists Association*.

Gella, Spandana, Paul Cook, and Timothy Baldwin. 2014. One sense per tweeter... and other lexical semantic tales of twitter. *EACL 2014*.

Geng, Xue, Hanwang Zhang, Jingwen Bian, and Tat-Seng Chua. 2015. Learning image and user features for recommendation in social networks. In *ICCV*, pages 4274–4282.

Geng, Xue, Hanwang Zhang, Zheng Song, Yang Yang, Huanbo Luan, and Tat-Seng Chua. 2014. One of a kind: User profiling by social curation. In *ACM MM*, pages 567–576. ACM.

Gottipati, Swapna, Minghui Qiu, Liu Yang, Feida Zhu, and Jing Jiang. 2013. Predicting user's political party using ideological stances. In *Social Informatics*.

Gray, Nicola J, Jonathan D Klein, Peter R Noyce, Tracy S Sesselberg, and Judith A Cantrill. 2005. Health information-seeking behaviour in adolescence: the place of the internet. *Social science & medicine*.

Greene, Jeremy A, Niteesh K Choudhry, Elaine Kilabuk, and William H Shrank. 2011. Online social networking by patients with diabetes: a qualitative evaluation of communication with facebook. *J Gen Intern Med*.

Groop, Leif. 2015. Genetics and neonatal diabetes: towards precision medicine. *The Lancet*.

Gu, Quanquan, Jie Zhou, and Chris HQ Ding. 2010. Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs. In *SDM*, pages 199–210. SIAM.

Guo, Ting, Xingquan Zhu, Jian Pei, and Chengqi Zhang. 2014. Snoc: streaming network node classification. In *ICDM*, pages 150–159. IEEE.

Gupta, Nitish and Sameer Singh. 2015. Collectively embedding multi-relational data for predicting user preferences. *arXiv preprint arXiv:1504.06165*.

Gupta, Sonal and Christopher D Manning. 2014. Spied: Stanford pattern-based information extraction and diagnostics. *ACL*.

Haas, Linda, Melinda Maryniuk, Joni Beck, Carla E Cox, Paulina Duker, Laura Edwards, Edwin B Fisher, Lenita Hanson, Daniel Kent, Leslie Kolb, et al. 2013. National standards for diabetes self-management education and support. *Diabetes care*.

Harvey, Morgan, Fabio Crestani, and Mark J Carman. 2013. Building user profiles from topic models for personalised search. In *CIKM*, pages 2309–2314. ACM.

Hawn, Carleen. 2009. Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health affairs*.

He, Danning, Simon C Mathews, Anthony N Kalloo, and Susan Hutfless. 2014a. Mining high-dimensional administrative claims data to predict early hospital readmissions. *JAMIA*.

He, Xiangnan, Min-Yen Kan, Peichu Xie, and Xiao Chen. 2014b. Comment-based multi-view clustering of web 2.0 items. In *WWW*.

He, Xiaofei, Deng Cai, and Partha Niyogi. 2005. Laplacian score for feature selection. In *NIPS*.

Hecht, Brent, Lichan Hong, Bongwon Suh, and Ed H Chi. 2011. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. ACM.

Hofman, Jake M and Chris H Wiggins. 2008. Bayesian approach to network modularity. *PRL*.

Hu, Frank B. 2011. Globalization of diabetes the role of diet, lifestyle, and genes. *Diabetes care*.

Hu, Xia and Huan Liu. 2012. Text analytics in social media. In *Mining text data*. Springer.

Hu, Xia, Jiliang Tang, and Huan Liu. 2014. Online social spammer detection. In *AAAI*.

Hu, Xia, Lei Tang, Jiliang Tang, and Huan Liu. 2013. Exploiting social relations for sentiment analysis in microblogging. In *WSDM*.

Jain, Anil K. 2010. Data clustering: 50 years beyond k-means. *Pattern Recogn. Lett.*

Jalali, Ali, Sujay Sanghavi, Chao Ruan, and Pradeep K Ravikumar. 2010. A dirty model for multi-task learning. In *NIPS*.

Jamison-Powell, Sue, Conor Linehan, Laura Daley, Andrew Garbett, and Shaun Lawson. 2012. I can't get no sleep: discussing# insomnia on twitter. In *SIGCHI*.

Jin, Xin, Fuzhen Zhuang, Sinno Jialin Pan, Changying Du, Ping Luo, and Qing He. 2015. Heterogeneous multi-task semantic feature learning for classification. In *CIKM*.

Joachims, Thorsten et al. 2003. Transductive learning via spectral graph partitioning. In *ICML*.

Jones, Rosie, Ravi Kumar, Bo Pang, and Andrew Tomkins. 2007. I know what you did last summer: query logs and user privacy. In *CIKM*, pages 909–914. ACM.

Kalyanam, Janani, Sumithra Velupillai, Son Doan, Mike Conway, and Gert Lanckriet. 2015. Facts and fabrications about ebola: A twitter based study. *arXiv preprint arXiv:1508.02079*.

Katon, Wayne and Mark D Sullivan. 1990. Depression and chronic medical illness. *J Clin Psychiatry*.

Kelman, Herbert C. 1958. Compliance, identification, and internalization: Three processes of attitude change. *Journal of conflict resolution*.

Kim, Seyoung and Eric P Xing. 2009. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS genetics*.

Kimura, Rui, Satoshi Oyama, Hiroyuki Toda, and Katsumi Tanaka. 2007. Creating personal histories from the web using namesake disambiguation and event extraction. In *Web Engineering*. Springer, pages 400–414.

Klein, Samuel, Nancy F Sheard, Xavier Pi-Sunyer, Anne Daly, Judith Wylie-Rosett, Karmeen Kulkarni, and Nathaniel G Clark. 2004. Weight management through lifestyle modification for the prevention and management of type 2 diabetes: Rationale and strategies a statement of the american diabetes association, the north american association for the study of obesity, and the american society for clinical nutrition. *Diabetes care*.

Korda, Holly and Zena Itani. 2013. Harnessing social media for health promotion and behavior change. *Health promotion practice*.

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

Kuh, Diana and Yoav Ben Shlomo. 2004. *A life course approach to chronic disease epidemiology*. Number 2. Oxford University Press.

Kumar, Abhishek and Hal Daume III. 2012. Learning task grouping and overlap in multi-task learning. *ICML*.

Kumar, Abhishek, Piyush Rai, and Hal Daume. 2011. Co-regularized multi-view spectral clustering. In *NIPS*.

Kumar, Mrinal, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury.

2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *ACM Conference on Hypertext & Social Media*.

Lappas, Theodoros, Kun Liu, and Evimaria Terzi. 2009. Finding a team of experts in social networks. In *KDD*, pages 467–476. ACM.

Lee, Daniel D and H Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *NIPS*.

Lee, Joy L, Matthew DeCamp, Mark Dredze, Margaret S Chisolm, and Zackary D Berger. 2014. What are health-related users tweeting? a qualitative content analysis of health-related users and their messages on twitter. *JMIR*.

Leskovec, Jure, Kevin J Lang, and Michael Mahoney. 2010. Empirical comparison of algorithms for network community detection. In *WWW*.

Li, Chenliang, Aixin Sun, and Anwitaman Datta. 2012. Twevent: segment-based event detection from tweets. In *CIKM*, pages 155–164. ACM.

Li, Jiwei and Claire Cardie. 2014. Timeline generation: Tracking individuals on twitter. In *WWW*.

Li, Jiwei, Alan Ritter, Claire Cardie, and Eduard Hovy. 2014. Major life event extraction from twitter based on congratulations/condolences speech acts. In *EMNLP*.

Li, Jundong, Xia Hu, Jiliang Tang, and Huan Liu. 2015. Unsupervised streaming feature selection in social media. In *CIKM*, pages 1041–1050. ACM.

Li, Rui, Chi Wang, and Kevin Chen-Chuan Chang. 2014. User profiling in an ego network: co-profiling attributes and relationships. In *WWW*, pages 819–830. ACM.

Li, Rui, Shengjie Wang, and Kevin Chen-Chuan Chang. 2012. Multiple location profiling for users and relationships from social network and content. *Proceedings of the VLDB Endowment*, 5(11):1603–1614.

Li, Rui, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. 2012a. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *KDD*, pages 1023–1031. ACM.

Li, Zechao, Yi Yang, Jing Liu, Xiaofang Zhou, and Hanqing Lu. 2012b. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*.

Lim, Stephen S, Theo Vos, Abraham D Flaxman, et al. 2013. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions. *The lancet*.

Lin, Chen, Jiang-Ming Yang, Rui Cai, Xin-Jing Wang, and Wei Wang. 2009. Simultaneously modeling semantics and structure of threaded discussions: a sparse coding approach and its applications. In *SIGIR*.

Lin, Huijie, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. 2014. User-level psychological stress detection from social media using deep neural network. In *ACM MM*.

Liu, Bing. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*.

Liu, Chuanren, Fei Wang, Jianying Hu, and Hui Xiong. 2015. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *KDD*.

Liu, Haifeng, Zhaohui Wu, Xuelong Li, Deng Cai, and Thomas S Huang. 2012. Constrained nonnegative matrix factorization for image representation. *PAMI*.

Liu, Jun, Shuiwang Ji, and Jieping Ye. 2009a. Multi-task feature learning via efficient l 2, 1-norm minimization. In *AUAI*.

Liu, Jun, Shuiwang Ji, and Jieping Ye. 2009b. Multi-task feature learning via efficient l 2, 1-norm minimization. In *UAI*, pages 339–348. AUAI Press.

Liu, Yunzhong, Yi Chen, Jiliang Tang, and Huan Liu. 2015. Context-aware experience extraction from online health forums. In *Healthcare Informatics (ICHI), 2015 International Conference on*, pages 42–47. IEEE.

Lorig, Kate, Philip L Ritter, Diana D Laurent, Kathryn Plant, Maurice Green, Valarie Blue Bird Jernigan, and Siobhan Case. 2010. Online diabetes self-management program a randomized study. *Diabetes care*.

Lu, Caimei, Xin Chen, and EK Park. 2009. Exploit the tripartite network of social tagging for web clustering. In *CIKM*.

Luxton, David D, Jennifer D June, and Jonathan M Fairall. 2012. Social media and suicide: a public health perspective. *American Journal of Public Health*.

Ma, Xiaoke, Lin Gao, Xuerong Yong, and Lidong Fu. 2010. Semi-supervised clustering algorithm for community structure detection in complex networks. *Physica A*.

MacLean, Diana, Sonal Gupta, Anna Lembke, Christopher Manning, and Jeffrey Heer. 2015. Forum77: An analysis of an online health forum dedicated to addiction recovery. In *CSCW*.

Maibach, Edward W and David Cotton. 1995. Moving people to behavior change: a staged social cognitive approach to message design.

Majumder, Anirban and Nisheeth Shrivastava. 2013. Know your personalization: learning topic level personalization in online services. In *WWW*, pages 873–884. International World Wide Web Conferences Steering Committee.

Malanda, Uriëll L, Sandra D Bot, and Giel Nijpels. 2013. Self-monitoring of blood glucose in noninsulin-using type 2 diabetic patients it is time to face the evidence. *Diabetes Care*.

Mathioudakis, Michael and Nick Koudas. 2010. Twittermonitor: trend detection over the twitter stream. In *SIGMOD*, pages 1155–1158. ACM.

McAuley, Julian, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *KDD*, pages 785–794. ACM.

McAuley, Julian, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*, pages 43–52. ACM.

Mejova, Yelena, Hamed Haddadi, Anastasios Noulas, and Ingmar Weber. 2015. # foodporn: Obesity patterns in culinary interactions. In *Proceedings of the 5th International Conference on Digital Health 2015*, pages 51–58. ACM.

Meladianos, Polykarpos, Giannis Nikolentzos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. 2015. Degeneracy-based real-time sub-event detection in twitter stream. In *ICWSM*.

Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL*.

Mirnezami, Reza, Jeremy Nicholson, and Ara Darzi. 2012. Preparing for precision medicine. *NEJM*.

Mislove, Alan, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. 2010. You are who you know: inferring user profiles in online social networks. In *WSDM*, pages 251–260. ACM.

Moens, Marie-Francine, Juanzi Li, and Tat-Seng Chua. 2014. *Mining user generated content*. CRC Press.

Moghaddam, Samaneh, Mohsen Jamali, and Martin Ester. 2012. Etf: extended

tensor factorization model for personalizing prediction of review helpfulness. In *WSDM*, pages 163–172. ACM.

Murray, Christopher JL and Alan D Lopez. 1996. Evidence-based health policy–lessons from the global burden of disease study. *Science*.

Namata, Galileo Mark, Brian Staats, Lise Getoor, and Ben Shneiderman. 2007. A dual-view approach to interactive network visualization. In *CIKM*.

Nesterov, Y. 2004. Introductory lectures on convex optimization: a basic course.

Newman, Mark EJ. 2006. Modularity and community structure in networks. *PNAS*.

Newman, Mark EJ and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E*.

Nie, Feiping, Heng Huang, Xiao Cai, and Chris H Ding. 2010. Efficient and robust feature selection via joint $l_{2/1}$-norms minimization. In *NIPS*.

Nie, Liqiang, Mohammad Akbari, Tao Li, and Tat-Seng Chua. 2014a. A joint local-global approach for medical terminology assignment. In *SIGIR*.

Nie, Liqiang, Tao Li, Mohammad Akbari, Jialie Shen, and Tat-Seng Chua. 2014b. Wenzher: Comprehensive vertical search for healthcare domain. In *SIGIR*.

Nie, Liqiang, Yi-Liang Zhao, Mohammad Akbari, Jialie Shen, and Tat-Seng Chua. 2015. Bridging the vocabulary gap between health seekers and healthcare knowledge. *TKDE*.

Nori, Nozomi, Hisashi Kashima, Kazuto Yamashita, Hiroshi Ikai, and Yuichi Imanaka. 2015. Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care. In *KDD*.

Nowson, Scott and Jon Oberlander. 2006. The identity of bloggers: Openness and gender in personal weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 163–167.

Obozinski, Guillaume, Ben Taskar, and Michael I Jordan. 2010. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*.

Omidvar-Tehrani, Behrooz, Sihem Amer-Yahia, and Alexandre Termier. 2015. Interactive user group analysis. In *CIKM*, pages 403–412. ACM.

Pan, Weike and Li Chen. 2013. Gbpr: Group preference based bayesian personalized ranking for one-class collaborative filtering. In *IJCAI*.

Pan, Yingwei, Ting Yao, Tao Mei, Houqiang Li, Chong-Wah Ngo, and Yong Rui. 2014. Click-through-based cross-view learning for image search. In *SIGIR*.

Papadopoulos, Symeon, Yiannis Kompatsiaris, and Athena Vakali. 2010. A graph-based clustering scheme for identifying related tags in folksonomies. In *Data Warehousing and Knowledge Discovery*. Springer, pages 65–76.

Park, Kunwoo, Ingmar Weber, Meeyoung Cha, and Chul Lee. 2015. Persistent sharing of fitness app status on twitter. *arXiv preprint arXiv:1510.04049*.

Pastors, Joyce Green, Hope Warshaw, Anne Daly, Marion Franz, and Karmeen Kulkarni. 2002. The evidence for the effectiveness of medical nutrition therapy in diabetes management. *Diabetes care*.

Paul, Michael, Mark Dredze, David Broniatowski, and Nicholas Generous. 2015. Worldwide influenza surveillance through twitter. In *AAAI*.

Paul, Michael J and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. *ICWSM*, 20:265–272.

Paul, Michael J and Mark Dredze. 2014. Discovering health topics in social media using topic models. *PLoS One*.

Peersman, Claudia, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Pennebaker, James W, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*.

Petrović, Saša, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics.

Phuvipadawat, Swit and Tsuyoshi Murata. 2010. Breaking news detection and tracking in twitter. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, pages 120–123. IEEE.

Powers, David Martin. 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Preoţiuc-Pietro, Daniel, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through twitter content. In *ACL*. The Association for Computational Linguistics.

Qi, Guo-Jun, Charu C Aggarwal, and Thomas Huang. 2012. Community detection with edge content in social media networks. In *ICDE*.

Qu, Yan and Jun Zhang. 2013. Trade area analysis using user generated mobile location data. In *WWW*. International World Wide Web Conferences Steering Committee.

Ramage, Daniel, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, pages 248–256. Association for Computational Linguistics.

Rao, Delip, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd interna-*

tional workshop on Search and mining user-generated contents, pages 37–44. ACM.

Ritter, Alan, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *KDD*.

Ritterman, Joshua, Miles Osborne, and Ewan Klein. 2009. Using prediction markets and twitter to predict a swine flu pandemic. In *1st international workshop on mining social media*, volume 9, pages 9–17. ac. uk/miles/papers/swine09. pdf (accessed 26 August 2015).

Robinson, Peter N. 2012. Deep phenotyping for precision medicine. *Human mutation*.

Rosasco, Lorenzo, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. 2004. Are loss functions all the same? *Neural Computation*.

Ruan, Yiye, David Fuhry, and Srinivasan Parthasarathy. 2013. Efficient community detection in large networks using content and links. In *WWW*.

Ruths, Derek, Jürgen Pfeffer, et al. 2014. Social media for large studies of behavior. *Science*.

Ruvolo, Paul and Eric Eaton. 2014. Online multi-task learning via sparse dictionary optimization. In *AAAI*.

Sayyadi, Hassan, Matthew Hurst, and Alexey Maykov. 2009. Event detection and tracking in social streams. In *ICWSM*.

Schwartz, H Andrew, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Schwartz, H ANDREW, MAARTEN Sap, MARGARET L Kern, JOHANNES C Eichstaedt, ADAM Kapelner, MEGHA Agrawal, EDUARDO Blanco, LUKASZ Dziurzynski, GREGORY Park, DAVID STILLWELL, et al. 2016. Predicting individual well-being through the language of social media. In *Biocomputing*.

Shalizi, Cosma Rohilla and Andrew C Thomas. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods Res.*

Shelley, Kristina J. 2005. Developing the american time use survey activity classification system. *Monthly Lab. Rev.*

Shen, Yelong and Ruoming Jin. 2012. Learning personal+ social latent factor model for social recommendation. In *KDD*. ACM.

Shi, Jianbo and Jitendra Malik. 2000. Normalized cuts and image segmentation. *PAMI*.

Shin, Donghyuk, Suleyman Cetintas, Kuang-Chih Lee, and Inderjit S Dhillon. 2015. Tumblr blog recommendation with boosted inductive matrix completion. In *CIKM*, pages 203–212. ACM.

Shoham, David A, Liping Tong, Peter J Lamberson, Amy H Auchincloss, Jun Zhang, Lara Dugas, Jay S Kaufman, Richard S Cooper, and Amy Luke. 2012. An actor-based model of social network influence on adolescent body size, screen time, and playing sports. *PloS one*.

Sohn, Sunghwan, Cheryl Clark, Scott R Halgrim, Sean P Murphy, Christopher G Chute, and Hongfang Liu. 2014. Medxn: an open source medication extraction and normalization tool for clinical text. *JAMIA*.

Song, Xuemeng, Zhao-Yan Ming, Liqiang Nie, Yi-Liang Zhao, and Tat-Seng Chua. 2016. Volunteerism tendency prediction via harvesting multiple social networks. *TOIS*.

Song, Xuemeng, Liqiang Nie, Luming Zhang, Mohammad Akbari, and Tat-Seng Chua. 2015a. Multiple social network learning and its application in volunteerism tendency prediction. In *SIGIR*.

Song, Xuemeng, Liqiang Nie, Luming Zhang, Maofu Liu, and Tat-Seng Chua.

2015b. Interest inference via structure-constrained multi-source multi-task learning. In *IJCAI*, pages 2371–2377.

Sun, Zhaonan, Fei Wang, and Jianying Hu. 2015. Linkage: An approach for comprehensive risk prediction for care management. In *KDD*.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*.

Tang, Duyu, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL*.

Tang, Jiliang, Huiji Gao, Xia Hu, and Huan Liu. 2013. Exploiting homophily effect for trust prediction. In *WSDM*, pages 53–62. ACM.

Tang, Jiliang and Huan Liu. 2012. Unsupervised feature selection for linked social media data. In *KDD*.

Tang, Jiliang, Xufei Wang, Huiji Gao, Xia Hu, and Huan Liu. 2012. Enriching short text representation in microblog for clustering. *Frontiers of Computer Science*.

Tang, Lei, Xufei Wang, and Huan Liu. 2011. Group profiling for understanding social structures. *TIST*.

Taylor, Shelley E and Jonathon D Brown. 1988. Illusion and well-being: a social psychological perspective on mental health. *Psychological bulletin*.

Teodoro, Rannie and Mor Naaman. 2013. Fitter with twitter: Understanding personal health and fitness activity in social media. In *ICWSM*.

Thelen, Michael and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *EMNLP*.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*.

Ver Steeg, Greg, Aram Galstyan, and Armen E Allahverdyan. 2011. Statistical mechanics of semi-supervised clustering in sparse graphs. *Journal of Statistical Mechanics: Theory and Experiment*.

Volkovs, Maksims N and Guang Wei Yu. 2015. Effective latent models for binary feedback in recommender systems. In *SIGIR*.

Von Luxburg, Ulrike. 2007. A tutorial on spectral clustering. *Statistics and computing*.

Vosecky, Jan, Dan Hong, and Vincent Y Shen. 2009. User identification across multiple social networks. In *Networked Digital Technologies*.

Wallach, Hanna M, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *ICML*.

Wan, Xiaojun, Jianfeng Gao, Mu Li, and Binggong Ding. 2005. Person resolution in person search results: Webhawk. In *CIKM*, pages 163–170. ACM.

Wang, Chi, Rajat Raina, David Fong, Ding Zhou, Jiawei Han, and Greg Badros. 2011. Learning relevance from heterogeneous social network and its application in online targeting. In *SIGIR*.

Wang, Fei, Namyoon Lee, Jianying Hu, Jimeng Sun, Shahram Ebadollahi, and Andrew F Laine. 2013. A framework for mining signatures from event sequences and its applications in healthcare data. *PAMI*.

Wang, Fei, Ping Zhang, Buyue Qian, Xiang Wang, and Ian Davidson. 2014. Clinical risk prediction with multilinear sparse logistic regression. In *KDD*.

Wang, Fei, Jiayu Zhou, and Jianying Hu. 2014. Densitytransfer: A data driven approach for imputing electronic health records. In *ICPR*.

Wang, Pengfei, Jiafeng Guo, and Yanyan Lan. 2014. Modeling retail transaction data for personalized shopping recommendation. In *CIKM*, pages 1979–1982. ACM.

Wang, Suhang, Jiliang Tang, and Huan Liu. 2015. Embedded unsupervised feature selection. In *AAAI*.

Wang, Xiang, Buyue Qian, and Ian Davidson. 2012. Labels vs. pairwise constraints: A unified view of label propagation and constrained spectral clustering. In *ICDM*.

Wang, Xiang, David Sontag, and Fei Wang. 2014. Unsupervised learning of disease progression models. In *KDD*.

Wang, Yi-Chia, Robert Kraut, and John M Levine. 2012. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In *CSCW*.

Wang, Zhu, Daqing Zhang, Xingshe Zhou, Dingqi Yang, Zhiyong Yu, and Zhiwen Yu. 2014. Discovering and profiling overlapping communities in location-based social networks. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, 44(4):499–509.

Weber, Ingmar and Palakorn Achananuparp. 2015. Insights from machine-learned diet success prediction. In *PSB*.

Weng, Jianshu and Bu-Sung Lee. 2011. Event detection in twitter. *ICWSM*, 11:401–408.

Weston, Jason, Ron J Weiss, and Hector Yee. 2013. Nonlinear latent factorization by embedding multiple user interests. In *RecSys*.

Wing, Christopher and Hui Yang. 2014. Fityou: integrating health profiles to real-time contextual suggestion. In *SIGIR*.

Wu, Lei, Linjun Yang, Nenghai Yu, and Xian-Sheng Hua. 2009. Learning to tag. In *WWW*.

Xu, Hua, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. 2010. Medex: a medication information extraction system for clinical narratives. *JAMIA*.

Xu, Linli, Aiqing Huang, Jianhui Chen, and Enhong Chen. 2015. Exploiting task-feature co-clusters in multi-task learning. In *AAAI*.

Xu, Tingyang, Jiangwen Sun, and Jinbo Bi. 2015. Longitudinal lasso: Jointly learning features and temporal contingency for outcome prediction. In *CIKM*.

Yan, Xiang and Ling Yan. 2006. Gender classification of weblog authors. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 228–230.

Yang, Liang, Xiaochun Cao, Di Jin, Xiao Wang, and Dan Meng. 2015. A unified semi-supervised community detection framework using latent space graph regularization. *CYB*.

Yang, Shuang-Hong, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng, and Hongyuan Zha. 2011. Like like alike: joint friendship and interest propagation in social networks. In *WWW*, pages 537–546. ACM.

Yang, Tianbao, Rong Jin, Yun Chi, and Shenghuo Zhu. 2009. Combining link and content for community detection: a discriminative approach. In *KDD*, pages 927–936. ACM.

Yang, Yang, Jia Jia, Shumei Zhang, Boya Wu, Qicong Chen, Juanzi Li, Chunxiao

Xing, and Jie Tang. 2014a. How do your friends on social media disclose your emotions? In *AAAI*.

Yang, Yuhao, Chao Lan, Xiaoli Li, Bo Luo, and Jun Huan. 2014b. Automatic social circle detection using multi-view clustering. In *CIKM*.

Yoshida, Minoru, Masaki Ikeda, Shingo Ono, Issei Sato, and Hiroshi Nakagawa. 2010. Person name disambiguation by bootstrapping. In *SIGIR*, pages 10–17. ACM.

Yu, Philip S and Jiawei Zhang. 2015. Mcd: Mutual clustering across multiple social networks. In *TBD*.

Yu, Shi, Bart De Moor, and Yves Moreau. 2009. Clustering by heterogeneous data fusion: framework and applications. In *NIPS workshop*.

Yuan, Lei, Yalin Wang, Paul M Thompson, Vaibhav A Narayan, and Jieping Ye. 2012. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *KDD*.

Zha, Hongyuan, Xiaofeng He, Chris Ding, Ming Gu, and Horst D Simon. 2001. Spectral relaxation for k-means clustering. In *NIPS*.

Zhang, Jiawei and Philip S Yu. 2015. Community detection for emerging networks. In *SDM*.

Zhang, Zhong-Yuan. 2013. Community structure detection in complex networks with partial background information. *EPL*.

Zhang, Zhong-Yuan, Kai-Di Sun, and Si-Qi Wang. 2013. Enhanced community structure detection in complex networks with partial background information. *Scientific reports*.

Zhao, Tong, Julian McAuley, and Irwin King. 2014. Leveraging social connections to improve personalized ranking for collaborative filtering. In *CIKM*.

Zhao, Tong, Julian McAuley, and Irwin King. 2015. Improving latent factor models via personalized feature projection for one class recommendation. In *CIKM*, pages 821–830. ACM.

Zhao, Xin Wayne, Yanwei Guo, Rui Yan, Yulan He, and Xiaoming Li. 2013a. Timeline generation with social attention. In *SIGIR*.

Zhao, Yi-Liang, Qiang Chen, Shuicheng Yan, Tat-Seng Chua, and Daqing Zhang. 2013b. Detecting profilable and overlapping communities with user-generated multimedia contents in lbsns. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 10(1):3.

Zhao, Zhe, Zhiyuan Cheng, Lichan Hong, and Ed H Chi. 2015. Improving user topic interest profiles by behavior factorization. In *WWW*.

Zhao, Zheng and Huan Liu. 2007. Spectral feature selection for supervised and unsupervised learning. In *ICML*.

Zhou, Deyu, Liangyu Chen, and Yulan He. 2015. An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *AAAI*.

Zhou, Jiayu, Jun Liu, Vaibhav A Narayan, Jieping Ye, et al. 2013. Modeling disease progression via multi-task learning. *NeuroImage*.

Zhou, Jiayu, Fei Wang, Jianying Hu, and Jieping Ye. 2014. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In *KDD*.

Zhou, Li, Genevieve B Melton, Simon Parsons, and George Hripcsak. 2006. A temporal constraint structure for extracting temporal information from clinical narrative. *Journal of biomedical informatics*.

Zhou, Wei, Clement Yu, Neil Smalheiser, Vetle Torvik, and Jie Hong. 2007. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *SIGIR*.

Zhou, Wenjun, Hongxia Jin, and Yan Liu. 2012. Community discovery and profiling with social messages. In *KDD*, pages 388–396. ACM.

Zhou, Yang and Ling Liu. 2013. Social influence based clustering of heterogeneous information networks. In *KDD*.

Zhuang, Jinfeng, Tao Mei, Steven CH Hoi, Xian-Sheng Hua, and Yongdong Zhang. 2015. Community discovery from social media by low-rank matrix recovery. *TIST*.