USING SOCIAL MEDIA WEBSITES TO SUPPORT SCENARIO-BASED DESIGN OF ASSISTIVE TECHNOLOGY

Xing Yu

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

January 2020

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____
Erin Brady, Ph.D., Chair

_____
Mathew Palakal, Ph.D.

August 23, 2019

_____
Davide Bolchini, Ph.D.

_____
Sunandan Chakraborty, Ph.D.

_____
Mohammad Hasan, Ph.D.

© 2020

Xing Yu

DEDICATION

To my family, friends, and my mentors.

ACKNOWLEDGEMENT

I would like to thank my parents and my wife Ran for their support through my Ph.D. training. All the work and the thesis would not be possible without their support and understanding.

I am truly grateful for my advisor Dr. Erin Brady, who has guided me through my Ph.D. training with passion, understanding, and support. I appreciate the freedom I had in pursuing this research project and the guidance I received to make the work robust and constantly improving.

I am also grateful for all my committee members, Dr. Mathew Palakal, Dr. Davide Bolchini, Dr. Sunandan Chakraborty, and Dr. Mohammad Hasan. Their passion and wisdom gave me inspirations and helped me to become a better researcher.

My sincere thanks to my friends, who I had the privilege to work with during my graduate study. I feel very lucky and appreciate all the support and the help I had along the way.

Finally, I would like to thank the IU School of Informatics and Computing and everyone within for this wonderful journey that I had in my lifetime.

Xing Yu

USING SOCIAL MEDIA WEBSITES TO SUPPORT SCENARIO-BASED DESIGN OF ASSISTIVE TECHNOLOGY

Having representative users, who have the targeted disability, in accessibility studies is vital to the validity of research findings. Although it is a widely accepted tenet in the HCI community, many barriers and difficulties make it very resource-demanding for accessibility researchers to recruit representative users. As a result, researchers recruit non-representative users, who do not have the targeted disability, instead of representative users in accessibility studies. Although such an approach has been widely justified, evidence showed that findings derived from non-representative users could be biased and even misleading. To address this problem, researchers have come up with different solutions such as building pools of users to recruit from. But still, the data is not widely available and needs a lot of effort and resource to build and maintain.

On the other hand, online social media websites have become popular in the last decade. Many online communities have emerged that allow online users to discuss health-related subjects, exchange useful information, or provide emotional support. A large amount of data accumulated in such online communities have gained attention from researchers in the healthcare domain. And many researches have been done based on data from social media websites to better understand health problems to improve the well-being of people.

Despite the increasing popularity, the value of data from social media websites for accessibility research remains untapped. Hence, my work aims to create methods that could extract valuable information from data collected on social media websites for

accessibility practitioners to support their design process. First, I investigate methods that enable researchers to effectively collect representative data from social media websites. More specifically, I look into machine learning approaches that could allow researchers to automatically identify online users who have disabilities (representative users). Second, I investigate methods that could extract useful information from user-generated free-text using techniques drawn from the information extraction domain. Last, I explore how such information should be visualized and presented for designers to support the scenario-based design process in accessibility studies.

Erin Brady, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

Chapter 1

Introduction

Having representative users in a scientific study is widely accepted as a tenet in the HCI community. However, finding and recruiting representative users is challenging. The problem is even more difficult to tackle when it comes to accessibility studies. Since the prospective participants can be hard to recruit because of their relative scarcity in the population. There might be accessibility problems with researchers' environments, which makes it difficult for participants to come to the lab setting. Plus, there could be other unexpected time and geographical limitations make this problem even harder. As described in existing literature (Sears & Hanson, 2012), representative users in accessibility study can be challenging to recruit. And the recruiting process can also be very time and resource demanding. A large number of existing accessibility studies involved information collected from non-representative users instead of representative users. There are two common scenarios. First is that non-representative users are asked to simulate specific disabilities when testing the new technologies. And the second is when studying activities that appeared to not be affected by the disability. Such approaches are usually explicitly or implicitly justified in the studies. However, a lot of literature (Sears, Karat, Oseitutu, Karimullah, & Feng, 2001; Heller, 1989; B. N. Walker & Mauney, 2010) revealed that studying non-representative users can lead to inaccurate and even wrong insights. To address this problem, researchers have created innovative approaches from different perspectives. Such as using remote settings to carry out evaluation experiments (Petrie, Hamilton, King, & Pavan, 2006) or creating pools of representative users to allow for easy recruitment (Dee & Hanson, 2014). However, gaining access to representative

users and acquire data generated by them in accessibility research is still difficult and expensive in general. A fast and cost-efficient way to gain access to data, as well as contact information to target groups would be beneficial to accessibility researchers and assistive technology designers. Especially when their projects are still in formative stages.

My work focuses on utilizing data available on social media websites to solve this problem. Social media websites have become a recognized data source for scientific research in recent years. On the one hand, using data collected from social media websites has many benefits. It provides abundant information with rich dimensions (textual data, geographic location information, media, etc.). The amount of data is scalable, and the data collection process is less intrusive and can be conducted asynchronously. All these advantages make social media a potentially valuable data source for studying participants that are hard to get access to using traditional methods. On the other hand, with the proliferation of social media, more and more people with health problems are discussing their conditions online to exchange information and provide support to each other (De Choudhury, Kiciman, Dredze, Coppersmith, & Kumar, 2016; M. Walker et al., 2015). There are many active online communities that contain valuable information, which is yet to be utilized. The anonymity nature of many social media websites also facilitates the online discussion to be more open than in lab settings(Ma, Hancock, & Naaman, 2016).In order to create a method to use data from social media websites, I need to answer three questions: 1)how to efficiently collect valid data from representative users on social media websites; 2) how to transform the data so it can be useful for designers of assistive technologies; 3) how can the data help

accessibility researchers to better understand representative users and create better support to help them. Based on the three questions, my work consists of three major parts. First, I create methods to automatically identify representative users on social media websites. Specifically, I use machine-learning methods to automatically label representative users in a graph that represents an online social network. Second, I develop methods to extract and transform the data generated by representative users for assistive technology designers. The common "language" that is used by HCI practitioners is "scenario". And I create a method based on entity-extraction to find important scenario related information from free text. Last, I look into how such information from representative users can help accessibility researchers better understand people with disabilities interact with information technologies. The main contribution of my work has two parts. From the methodological perspective, I develop a machine-learning method that can identify a new attribute, whether a user has a disability or not, of online users. I also, develop methods to extract useful information from data collected from social media website for researchers and designers. From a theoretical perspective, I look into how people with disabilities use social network online. And the findings would help accessibility researcher better understand the value of social media websites to disabled people. To my best knowledge, little work has been done to explore the value of social media websites in the accessibility domain. And none have tried to use data collected from social media websites to help practitioners in the accessibility domain to make better design decisions.

1.1 Structure of the Dissertation

The dissertation is organized as follows. In Chapter 1, I introduce the motivation of my work, which is trying to use social media as a data source for data collection and participants recruiting for researchers/designers in the accessibility domain. In Chapter 2, I formally raise the research questions that I try to answer. In Chapter3, I review the existing literature that is related to my work. More specifically, I review existing work that focuses on the homophily theory, online user classifications, and data visualization.

In Chapter 4, I introduce the phase one study of my dissertation, in which I try to explore potentially useful features to classify online users who have disabilities. I explored potential features from three categories: personal interests, psychological traits, and online community features. I used random forest and LASSO to construct classification models. The results showed that community-based features are the most useful in the classification task.

Based on the findings from Chapter 4, I proposed a new model in Chapter 5 to carry out the classification. The new model is a co-training model containing a varied version of label propagation model that works on a bipartite graph. I experimented the model with a dataset that consists of 6 different classes and the results showed improvements over baseline methods. Based on the new model that I created, I could identify online representative users and collect their user-generated data.

In Chapter 6, I introduce a new data visualization tool that I created to present the online user-generated data. The visualization tool is based on the PACT analysis, which is used in the scenario-based design. I created a pipeline of models to power the tool,

developed a frontend interface, and carried out a pilot study to evaluate the tool. Finally, in Chapter 7, I present the conclusion, implications, and future work of my research.

Chapter 2

Research Questions

This dissertation addresses the problem of using social media websites as a source for data collecting and participants recruiting in existing accessibility studies by developing methods to automatically identify online users with disabilities and extract useful information from online posts. I will address the following research questions (RQs) throughout my work.

RQ1: How to efficiently collect representative data from social media websites?

The data that exists in online communities related to disabilities share the same problem of unrepresentativeness. Healthcare or disabilities related communities on social media websites consist of both representative users who have disabilities and non-representative users who are close to representative users such as caregivers, family members, and health practitioners. Hence, the data collected from such communities can also be categorized into different classes by the types of its creators.

For accessibility researchers, collecting and analyzing the data generated by the stakeholders who have disabilities is most important. To achieve that, a filtering process is necessary. However, the sheer size of the data available on social media websites can make manual filtering prohibitive. In my work, I look into developing an automatic method to identify representative users and collecting data generated by them. Throughout the process, I answer the following sub-questions:

4.0 What are the differences between the representative users and non-representative users on social media websites?

5.0 How to automatically classify representative users and non-representative users?

RQ2: How to extract valuable information from social media websites for designers?

Data collected from social media websites can be of poor quality. More importantly, such data may not be in the desirable form to inform designers of assistive technologies.

Scenario-based design is an approach that is widely applied in the HCI community. A scenario is a set of one or more events in which one or more actors try to accomplish specific goal(s) in a given context. And itis the building block of scenario-based design, as scenarios can be used in multiple phases in a design circle. For example, researchers can use scenarios to explore the problem spaces by stimulating conversations in interviews or give scenarios to participants to explore potential problems of design in the evaluation process.

After collecting data generated by representative users, I look into how I can extract useful information from the data that consists of mostly natural language. I follow the PACT format of scenario analysis and create models to extract person, activity, context, and technologies related terms from freetext.

RQ3: How can the extracted information be visualized to help accessibility researchers and designers?

Data from social media can be overwhelming even after filtering. After extracting the information, how can such information be organized and visualized to help researchers and designers? I will exam the following question:

6.0 How should I organize and visualized the online textual data?

Chapter 3

Related Work

In this section, I provide background information for my dissertation. First, I introduce the definition of representative users and non-representative users in accessibility research, which are the most important concepts throughout this work. Next, I review the current work that is related to healthcare using social media websites. The information helps to explain the gap that I am trying to address in my work. After that, I review important theories and methods from which I draw inspirations to develop my own method to collect, extract, and present data generated by representative users on social media websites.

3.1 Definitions of Representative and Non-representative users

The concept of representative users in this dissertation is drawn from Andrew's work (Sears & Hanson, 2012), in which representative users refer to the population that is expected to have a unique set of disabilities, which may affect how they interact with information technologies. And non-representative users refer to the population that does not represent the intended users of the information technologies. A quick example would be that in a study of evaluating a screen reader, representative users are people who are visually impaired while non-representative users are people who could see clearly. Since different designs of information technologies have different intended users, I generalize the definition of representative users to people who have at least one type of disability in my dissertation. And non-representative users refer to people who don't have that disability.

3.2 Social Media in Health-Related Research

Social media websites such as Facebook and Twitter have become a type of popular information technology for both patients and health professionals in the past decade. According to the survey result (Antheunis, Tates, & Nieboer, 2013), patients use social media to increase knowledge, communicate with doctors, gain social support, and exchange advice for self-care. On the other hand, medical professionals not only use social media to communicate with patients but also for communication with colleagues and marketing. The result of the aforementioned phenomena is a large amount of data that is publicly available for researchers.

Existing literature that use social media data to study health-related problems can be largely divided into two categories. The first one is descriptive, in which researchers use a large amount of data to conduct empirical studies of online users who have medical problems. A popular topic in this category is studying online users who are visually impaired over social media websites (Morris et al., 2016; Wu & Adamic, 2014; Voykinska, Azenkot, Wu, & Leshed, 2016). In these studies, researchers use social media data to apply multiple analysis methods to understand behavioral patterns and social network dynamics of users. The second category is predictive, in which researchers use social media data to predict health problems. Given the advancement of machine-learning techniques and increasing amount of data, some predictive work try to address many problems such as mental health prevention(De Choudhury et al., 2016; De Choudhury, Gamon, Counts, & Horvitz, 2013), public health outbreak prediction (Schmidt, 2012; Paul & Dredze, 2011), etc.

There are two common issues that exist in current work. First, no matter what the type of work is, data filtration is a common issue that researchers need to deal with when studying health-related problems since itis necessary to identify a subgroup of users with the target health issue. Existing studies have used methods such as crowd-sourcing (De Choudhury et al., 2013) or information from a client software (Wu & Adamic,2014). The aforementioned method all applies to the specific context. And little work has been done to develop an efficient data filtration method that can be generalized. Second, despite many works have been done about health-related problems on social media, little is focusing on disability. More importantly, there lacks a link to convert the data from social media into insight for developing information technologies for people with disabilities.

3.3 Homophily

In this section, I review the homophily theory, which is important in model designing in late sections.

People's personal social networks are shaped by homophily (McPherson, Smith-Lovin, & Cook, 2001). As the homophily principle suggests, social ties between two similar individuals have higher chances to occur than dissimilar persons. Also, ties between dissimilar people are less stable and likely to dissolve at a higher rate.

The source of homophily includes but not limits to space, organization, occupation, cognitive processes, and family ties. Plus, the pattern of homophily tends to get stronger when there are multiple relationships between two individuals (Fischer, 1982), which suggests a cascading effect of sources of homophily.

The study of homophily originated with people's offline social networks and gradually moved onto online social networks. The same principle was found also

applicable to personal social networks over the Internet. In a study of Twitter (Kwak, Lee, Park, & Moon, 2010), authors found homophily among users who have reciprocated relationships by analyzing 106 million tweets. Also, there is research showing online users who follow reciprocally share topical interests (Weng, Lim, Jiang, & He, 2010). Not only was homophily applied to study online social ties among users, it is also used to model online phenomena. In (Aral, Muchnik, &Sundararajan, 2009), authors statistically modeled behavioral contagion in dynamic networks and found that homophily explains more than half of the perceived behavioral diffusion. One very important application of homophily in online social networks is to predict potential relationships. For example, authors in (Aiello etal., 2012) found proof of homophily in analyzing social and semantic features in three online social networks and used them to predict potential social links.

3.3.1 Classification based on Homophily

As aforementioned, sources of homophily do not limit to traditional ties such as spatial similarities. More findings emerged in studies of social media websites suggest such a pattern may exist among online users who share similar disabilities and illness. Authors in (Wu & Adamic, 2014) found that visually impaired online users have more friends who share the same disabilities in the online communities that they belong to. The strong evidence of homophily is one key inspiration to the assumption that I develop my methods in my work. I assume that people's online social network is largely shaped by similarity rather than by special vicinity, which has a huge impact on people's offline social network. Hence, by understanding how people with disabilities formed their online social network, I can use selected dimensions of homophily to identify users who share the same disabilities. A key challenge with applying the homophily theory for

classification is the complexity of similarities. As the theory suggests, a person's social network is a complexity since it has different layers that each is based on different levels of different dimensions of homophily. An intuitive example is that an online user may have one friend who shares the same disability and another friend who share the interests in sports cars. Hence, it is imperative to find the important dimensions in order to apply homophily theory into practice, which is online user classification in my work.

3.3.2 Dimensions of Homophily

Existing studies regarding people with disabilities have given insights of several promising dimensions among people who share same disabilities.

Cognitive Process

Cognitive process has been found to be effective in identifying online users with disabilities. Authors in (De Choudhury et al., 2016) studied how cognitive process through the measuring of function words usage, which can be used to detect mental health problem on social media websites. And the cause of such a phenomenon can be explained by the fact that people are experiencing emotional or physical pain tend to have different cognitive process (Rude, Gortner, & Pennebaker, 2004).

Similar Interests

Similar interests among users as an important factor in the creation of online communities has been explored and studied in general (Crandall, Cosley, Huttenlocher, Kleinberg, & Suri, 2008). Hence, I assume it is also an important factor that connects online users who have disabilities. Plus, given the nature of the communities, it is possible to assume that online users that participate in healthcare-related communities would share certain interests that differ from other communities.

Persistence of Online Interactions

The persistence of online interactions between two users may not fit in perfectly as a homophily dimension since it stems from certain similarities between two online users. However, as the principle suggests, the persistence of interactions contributes to the similarity of two users. In other words, the more frequent two users interact with each other, the more similar they become. According to our preliminary study (Yu &Brady, 2017), in which I built an online social network based purely on the frequency of online interactions among users, I found that features extracted from such a social network were very beneficial in identifying people who share the same disability.

The three dimensions that I have reviewed here were proved to be useful in our research. They are certainly not exhaustive in any sense. But I will test to what extent they can be used to carry out our classification task.

3.4 Methods for Modeling Online Communities

In this section, I review existing methods that are important to developing my method to identify representative users and model their online social networks.

3.4.1 General Online User Classification

Online users classification refers to existing studies that use the combination of knowledge from social science and computer science to identify online users' demographic information such as gender, age, region, and even political orientations to improve personalizing, marketing, and legal investigation (Schler, Koppel, Argamon, & Pennebaker, 2006; Argamon, Koppel, Pennebaker, & Schler, 2007; Rao, Yarowsky, Shreevats, &Gupta, 2010; Burger, Henderson, Kim, & Zarrella, 2011). Many approaches have been proven to be effective in discovering subtle differences in the content

generated by different groups of online users. Typically, the approaches focus on analyzing the textual content using a predefined dictionary such as Linguistic Inquiry and Word Count (LIWC) analysis, which is widely applied to conduct the psychometric analysis of language (Tausczik & Pennebaker, 2010), or statistic language models like Latent Dirichlet Allocation (Hoffman, Bach,& Blei, 2010).

3.4.2 Label Propagation

The text mining methods in the studies above have limits when dealing with identifying special groups of users. One major reason is that, even in social media, the number of people with disabilities is relatively small. Hence, the data will be severely unbalanced, which will impact the classification results. To remedy this problem, I will develop my method based on several important frameworks. Given the existing evidence of homophily, I introduce label propagation, which is a suitable semi-supervised learning algorithm based on regional smoothness on graphs. Label propagation algorithms were created to cope with the rapid growth of online multimedia content. A label propagation algorithm is typically a semi-automatic annotation process that labels a large number of unlabeled data points by using a set of labeled data points. The whole process is semi-supervised since it labels unlabeled data points using both labeled and unlabeled instances. And it is highly effective in labeling tasks where the manual annotation is prohibitive (Zoidi, Fotiadou, Nikolaidis, & Pitas, 2015).

The learning task of a label propagation algorithm is to spread labels from labeled instances to the unlabeled instances. A typical definition of a label propagation algorithm is as follow. Let's define the data set to be $X = \{x_1, \dots, x_n, x_{n+1}, \dots, x_N\}$ without loss of generalizability. The set $X_{label} = \{x_1, \dots, x_n\}$ denotes the n_l labeled instances while the

set $X_{unlabel} = \{x_{n+1}, \dots, x_N\}$ denotes the n_u unlabeled instances. And $n_l + n_u = N$. Apart from the data instances, there is also a vector $y = \{y_1, \dots, y_{n_l}, 0, \dots, 0\}$ that each $y_i \in L$, where $L = \{l_1, \dots, l_K\}$ is a set of all possible labels for the data points in X. In computation, there usually is a matrix $Y \in R^{N \times K}$ to represent the label information of X. If $y_{ij} = 1$, then the ith entry has the jth label in L. The propagation function is defined as in equation 3.1. In this equation, T denotes the adjacency matrix of a graph. C denotes the class assignments for each node in the graph. The superscripts l and u denote the labeled and unlabeled set in the graph. This propagation is iterated until convergence through the execution of the label propagation algorithm.

$$\begin{bmatrix} C^l \\ C^u \end{bmatrix} \coloneqq \begin{bmatrix} T^{ll} & T^{lu} \\ T^{ul} & T^{uu} \end{bmatrix} \cdot \begin{bmatrix} C^l \\ C^u \end{bmatrix} \quad (3.1)$$

As aforementioned, the label propagation will take the label information in X_{label} and spread it to $X_{unlabel}$ to finish the annotation process. And it takes two principles into considerations while doing so. First, the process will keep the original label information consistent, that is to keep the $y_{label} = \{y_1, \dots, y_n\}$ the same as before the propagation process. Second, the process should maximize local smoothness, which indicates that two data instances should have the same labels if they are similar to each other. The similarity between two data points is usually measured using a distance function. And thus, the label propagation algorithms are typically carried out on a graph that consists of node and edge, where each node represents a data point and each edge has a weight that represents the similarity between the two nodes it connects.

A typically label propagation process include two components. The first is a graph for the propagation. Many different methods have been proposed to construct graphs that facilitate regional smoothness. The most common way is to construct a

complete graph, in which each node is connected to the rest of the nodes by an edge, using a distance function such as the Radial Basis Function (RBF):

$$W_{ij} = exp\left(-\frac{\left|\left|x_i - x_j\right|\right|^2}{2\sigma^2}\right)$$

However, the choice of the parameter σ could greatly impact the performance of the algorithm. to address the problem, a local reconstruction method was introduced (Wang, Zhang, Shen, & Wang, 2006), in which each node is reconstructed using the linear combinations of its k nearest neighbors. And each node is only connected to the k nearest neighbors in the final graph. Hence a sparse graph is constructed for propagation. There also exist other methods to construct a sparse graph instead of a complete one. And most of them are based on selecting the k nearest neighbors (Talukdar, 2009; Satuluri, n.d.). The second component is label inference on the graph to spread label information from labeled nodes to unlabeled nodes. Typical methods include iterative algorithms, which gradually spread labeled information over the graph until it reaches a stationary state (Zhu & Ghahramani, 2002). Another common approach is using random walk on the graph to propagate label information (Szummer & Jaakkola, 2001; Baluja etal., 2008). An unlabeled node gets information from all other nodes in the graph. And the amount of label information it receives from each node is based on the commute time it has from that node. Hence, each node has a probability distribution over all possible labels.

3.4.3 Graph Representations

The key to the success of label propagation is the construction of the graph rather than the propagating process. And label propagation algorithms are not confined to a single graph. Since entities in real life can be represented in different ways, another form

of label propagation algorithm is to extend the process to a multi-representation scenario. For example, when representing online videos, one can use tags or user clicks to create two graphs that each is based on one of the similarities of the two metrics. And for each graph, an independent label propagation process can be carried out. And the results will be merged using certain methods. Existing studies focus on how to fuse the different representations to achieve the best results. The fusion process could take place before the label propagation, which is called early fusion, or after the label propagation, which is called late fusion. Generally speaking, late fusion outperforms early fusion since vectors with high dimensionless tend to be similar to each other (Snoek, Worring, & Smeulders, 2005). Furthermore, late fusion methods can be categorized into two types: linear fusion and sequential fusion. Existing studies show evidence that sequential fusion performs better than linear fusion (Tong, He, Li, Zhang, & Ma, 2005).

Label propagation techniques are appropriate to my task at hand given our assumption of homophily among online users with same disabilities. My key assumption is that, since homophily has many dimensions, people who share the same disability should be similar in several key dimensions. By allowing the labels to propagate on these dimensions, I can find people who share the same disability.

3.4.4 Deep Models in Node Classification

One area that has gain popularity in node classification is known as node embedding. One of such classic models is introduced in the node2vec (Grover & Leskovec, 2016a) mode. In this type of models, nodes in a graph are embedded in a low dimensional space based on their similarities. The overall target function of the model is equation 3.2. f denotes the mapping function that maps a node u into a vector. $N(u)$

denotes the neighbors of node u. The overall target is to maximize the log-probability of observing a network neighborhood for a node. The model is proven to be highly scalable and robust. However, the performance depends on how the neighbors $N(u)$ are defined in a specific task.

$$\underset{f}{argmax} \sum log\ Pr\big(N(u)\big|f(u)\big) \quad (3.2)$$

3.5 Data Visualization

In this section, I will review existing work that is relevant to creating my data visualization tool. I will focus on two aspects, which are data visualization strategies and the machine learning models that power the tool.

3.5.1 Scenarios-Based Design and Visualization

Scenario-based design is an approach just like other user-centered techniques. It helps to change the focus from defining system functions to how users interact with the system to accomplish goals (Benyon &Macaulay, 2002). Generally speaking, scenarios are stories that have a sequence of actions and events toward an outcome. A scenario typically consists of a setting, one or several actors, and objects that actors manipulate (Rosson & Carroll, 2009). The approach is effective for two main reasons. First, scenarios are versatile. Scenarios can be used in exploring problem spaces, facilitating conversations during interviews, helping set the context for evaluation of prototypes and so on. The second reason is that unlike other approaches that require formal analysis of human behavior, using scenarios is a fast and agile way to explore problems and possibilities.

Information Extraction

Information extraction (IE) is a research area that falls into the category of natural language processing (NLP). And it was originally initiated for the purpose of extracting

useful information for military purpose (Cowie &Lehnert, 1996). Information extraction refers to the tasks that identify the factual terms from freetext that is unstructured and turn it into structured data. Such factual terms include but not limited to names, locations, organization, and relationships. IE is different from information retrieval, which typically returns a ranked list of relevant documents. It focuses on understanding and extracting target information from a corpus. And it is also different from full-text understanding that tries to understand the entire semantic meaning of a text in natural language. Since it only focuses on extracting information that belongs to a predefined domain.

Types of Information Extraction

Based on what to extract, IE has four major types of tasks. The first one is called named entity recognition (NER). NER refers to the problem of identifying predefined name entities. The named entities could be persons, organizations, locations, and so on. The second type is co-reference recognition (CO), which aims at identifying multiple co-referring of the same entity in the text. For example, in the sentence "Amy came late, and she seemed angry.", "Amy" and "she" both refer to the person. The third type of IE is relation extraction (RE), which classifies the predefined relationships among entities. And the fourth type of IE is event extraction (EE), which tries to identify predefined events in the unstructured text. EE is the most difficult one among the four tasks since it tries to identify multiple entities and their relationships.

Existing Information Extraction Methods

An IE system has two major parts. The first part is domain-independent that use common techniques in natural language processing. Components in this part typically include a tokenizer, a stemmer, and sentence boundary detector. The second part is

domain-dependent, which handles the IE tasks that focus on a specific domain such as extracting medical entities or terrorist attacks. The key requirement for the domain-dependent part is domain knowledge. A knowledge base is required to conduct information extraction, and the process of creating a knowledge base is called knowledge engineering (KE) that requires heavy manual effort (Piskorski & Yangarber, 2013). In recent years, trainable IE systems emerged due to the proliferation of machine-learning. Supervised machine-learning methods such as hidden Markov models and conditional random field (Bikel, Miller, Schwartz, & Weischedel, 1997; Riloff et al.,1993) have been applied to solve IE problems. With the trainable IE system, knowledge engineering is replaced with the effort of feature engineering and text annotation. The requirement for manual efforts is less than KE, but it is still a complex task and resource demanding. To alleviate the problem, methods such as active learning and bootstrapping have been used in create annotations to improve efficiency. With the proliferation of big data and deep learning models, new techniques have been introduced to the field of IE. One famous framework proposed by Collobert et al. is to use a recurrent neural network (RNN)for named entity recognition (NER) (Collobert et al., 2011). A typical RNN is depicted in Figure 3.1. x_t denotes the input word. e_t is an embedding layer, which passes the embedding vector into h_t. The hiddenstatushtis calculated based previous output $h_{t-1}$and the input of the current step e_t. And the value is passed onto the next step. When this framework is adopted in a NER task, it becomes a sequence-to-sequence mode, which means there is one output for each input as each step. The h_t is passed through a softmax

function, which is denoted as σ. And the output is corresponding to a label for this word, which is denoted as o_t.

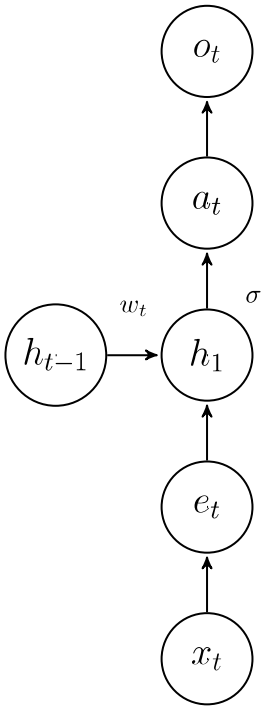

Figure 3.1: An RNN framework for NER

Information Extraction and PACT Analysis

Traditionally, scenarios can be derived from user interviews, focus groups, ethnographic studies, and soon. A typical structure for a scenario contains people, activities, contexts, technologies (PACT) (Benyon &Macaulay, 2002). And analysts have to look through the data to identify the PACT object.

It is obvious to see the similarity between the PACT analysis and IE tasks from the example in Table 3.1.A good scenario structure should have PACT objects. Hence, if I can extract PACT objects from freetext, then the free text can constitute a potential scenario. Although IE in social media has been a popular research topic in the last

decade, little work tries to use data from social media to extract knowledge to support HCI practitioners.

Scenario
John is a freshman at his 20's. He lives on campus and goes to classes on bicycles every day. Recently, John is looking for an apartment off campus. He goes to a rental list website where people post rental information to find an ideal one. The website has a map that shows the location of the available apartments. John wants one close to campus. Hence, he clicks on the nearby ones to check the facilities and prices. After a while, he found one that looks perfect. He sent a message via the website to schedule a meeting with the leasing agent.
PACT analysis
People: John, college freshman, 20's, healthy Activities: Using a leasing website to find a rental apartment for himself Context: Currently living on campus, want his own place Technology: A website with leasing information, available leases are presented on a map, can send messages via the website

Table 3.1: An example of PACT analysis on a fictitious scenario.

Divide and Conquer

Reorganizing a large dataset through breaking into parts and then combining later can also produce insights that might have been unreachable without portioning. One popular strategy for handling such large dataset is "divide-and-conquer." Roberts et al describe divide-and-conquer as a strategy that presents a portion of a challenge to each user and allowing the users to work in parallel to reach the insight that would have been hidden otherwise (Roberts, Lyons, Cafaro, & Eydt, 2014). Similarly, Roschelle and

Teasley emphasizes divide and conquer strategy as means for collaboration in "joint problem space" that share goals, problem state, possible activities, and other relevant information (Roschelle & Teasley, 1995). They differentiate the strategy from cooperation, which refers to having each person working on their portion of the problem without interacting. At the same time, as an algorithm, divide-and-conquer means dividing the dataset into small parts and then combining the insights generated in each part into synthesis.

Using Metadata as Filters

Big datasets are usually not only large in size but also rich in dimensions. In order to truly understand a large dataset, a tool should be able to scale in terms of the number of records (Robertson, Ebert, Eick, Keim, &Joy, 2009). However, how to create a standard format to scale the datasets is a big problem. Light et al (Light, Polley, & B¨orner, 2014) pointed out that when studying a social system, it is necessary to conduct studies from different aspects to truly understand the system. For example, researchers usually conduct longitudinal studies, geo-spatial studies swell as cross-sectional studies to complement each other. Based on the idea, they developed a Sci2 tool that extracts time, geo-spatial location, topic, and network information to provide answers of when, where, what, and with whom to researchers. In other words, the metadata extraction strategy was used to support the meta-level studies of big data in their study.

Given the proliferation of various tools like Hadoop (Shvachko, Kuang, Radia, & Chansler, 2010) and Spark (Shoro & Soomro, 2015) that could handle large quantitative data, I want to explore the possibility of a tool that could support the exploration of large qualitative datasets. To uncover the underlying patterns of datasets, there are generally

four dimensions to explore in data processing (Light et al., 2014). In my study, I want to focus on processing qualitative data from the topical dimension to help data exploration.

3.5.2 Visualization of Social Media Data

There has been a number of researches that utilized qualitative data from social media—mostly Twitter—for various purposes. First, some researchers used such data to detect the occurrence of certain events. Sakakiet al built a probabilistic model for detecting an occurrence of an earthquake and its location based on data retrieved from real-time tweets posted on Twitter (Sakaki, Okazaki, & Matsuo, 2010). The team created an algorithm to separate messages that are describing real earthquake from the messages that are merely referring to related events–an earthquake conference–. They evaluated tweets based on the number of words used, the meaning of the words, and the position of the earthquake-related keywords inside the tweets. Then they designated the users who posted relevant tweets as "sensors" for detecting an earthquake. Likewise, Corley et al analyzed blog posts that contain words related to influenza, which showed a correlation to the actual outbreak of influenza (Corley, Cook, Mikler, & Singh, 2010).

Also, other researchers analyzed tweets to make predictions about the events yet to come. De Choudhury et al suggested the possibility of using tweets to identify individuals that were likely to experience depression (De Choudhury et al., 2013). Along with gathering quantitative data such as a number of posts per day, they assessed qualitative data such as the linguistic property and the emotion expressed through texts. They also looked for specific terms that were related to depression. From their research, they were able to create a method that can predict depression with 70% accuracy. Furthermore, Bollen et al conducted an analysis of text contents of tweets to explore the

correlation between the moods expressed in those tweets and the changes in stock price (Bollen, Mao, & Zeng, 2011). From their analysis, they discovered that the calmness of public, which is expressed in tweets, showed relation to changes in DJIA (Dow Jones Industrial Average). They also found that they could build a more accurate prediction model of stock market change by including the degree of expressed calmness in the model. In addition, the method to identify such sentiments has also been the topic of research for many researchers. Wilson et al described a method for dividing sentiments into neutral and polar, and then identifying the extremity of polar sentiments (Wilson, Wiebe, &Hoffmann, 2005). Godbole et al also developed a system that assigned a score according to the level of positive and negative emotion expressed in the tweets to identify sentiments of a large number of users toward certain persons or events (Godbole, Srinivasaiah, & Skiena, 2007).

Moreover, qualitative data from social media could be used as sources of business insights. Analyzing postings from Twitter and Facebook on major pizza makers, He et al suggested utilizing data to monitor the rival companies, reveal patterns and trends regarding the consumers, and evaluate the effectiveness of promotions (He, Zha, & Li, 2013).

In addition, Naaman focused on retrieving non-textual data such as images and videos and making use of them to provide better web search experience (Naaman, 2012). He created Flickr Landmarks forgathering images from Flickr and Concert Sync for gathering videos, although both applications were not fully implemented. Flickr Landmarks utilized metadata on photos uploaded to Flickr to create labels that can represent landmarks in different countries. Then a user could see what sightseeing sites

were present in a certain area and select one of them to look at the photos related to the site. Concert Sync utilized metadata on videos uploaded to YouTube to make it easier for the user to search for a particular moment at the particular musical event.

Furthermore, there are researches that focused on visualizing social media data. Dou et al (Dou, Wang,Skau, Ribarsky, & Zhou, 2012) developed an interactive visualizing system LeadLine that can identify important event on its own and enable the user to explore the information related to the event. After LeadLine detected events from Twitter and news media, the user could choose a specific time, topic words related to the events, and geo-location to reshape the graph and discover underlying relationships and insights. Dork etal (D¨ork, Gruen, Williamson, & Carpendale, 2010) also proposed a medium called Visual Backchannel, which visualized the reactions toward a certain event in the form of a graph along with the photographs associated with the event. Using textual data from Twitter, they created an interactive "topic stream" that showed the popular topics related to the event that changed its shape due to the popularity of each word and appearance of new words. They also retrieved images posted in relation to the event and displayed them in the system.

At the same time, some researchers identified challenges in analyzing qualitative data from social media. Kleinberg discussed two concerns, which are 1) difficulty of building an explanation for phenomena that include multiple domains as they include different types of population, and 2) difficulty of preserving privacy which is not fully secured even through the existing measure for anonymization (Kleinberg, 2007). In addition, Maynard et al discussed specific challenges involving mining of text data, such as the tendency of data to contain sarcasm, more linguistic variation and lenient use of

grammar, or meaning that is only revealed in certain context (Maynard, Bontcheva, & Rout, 2012). While these challenges are present, Kouloumpis etal discovered that combining micro-blogging features such as emoticons and words in capital letters with the existence of negative, positive, or neutral words could be effective in assessing the sentiment in twitter postings (Kouloumpis, Wilson, & Moore, 2011).

3.5.3 Prototyping and Evaluation

Two most important parts in studies of data visualization are prototyping/implementation and evaluation. The two components are highly dependent on each other. A prototype/implementation is a necessary condition for evaluation while a rigorous evaluation would help verify the validity of the underlying design principles of the prototype/implementations. In practice, a good prototype/implementation with a reliable evaluation can be quite challenging in designing a research project (Plaisant, 2004). For example, unreliable results may be generated when an evaluation is carried out on the interface of a low-fidelity prototype (Chittaro & Dal Cin,2002). And, the results might incorporate additional variables like distractions that need to be taken into consideration (Chittaro & De Marco, 2004). In this section, I review the prototypes and implementations in existing research as well as the methods employed to carry out the evaluation.

The prototyping/implementation of data visualization tools in existing studies utilize different approaches, which vary in scale, infrastructure, and purpose. For example, Light et al (Light et al., 2014) created a database-tool that downloads, parses, mines, and visualizes data downloaded from online databases to support science studies. The tool they created was fully functional in order to see how the underlying algorithms

performed in visualizing large datasets. On the other hand, researchers in some studies focused more on the interaction part. Theophanis and his colleagues created SketchSliders (Tsandilas, Bezerianos, & Jacob,2015)—a prototype based on Wizard of OZ setting that users can use sketches on mobile devices to create different data visualizations on a wall display.

As aforementioned, another essential component in data visualization studies is evaluation. And both qualitative and quantitative research methods could be found in existing studies. In quantitative evaluation, a large range of different metrics, both behavioral and computational, were covered. The evaluation of visualization tools that were designed especially for the processing of big data focused more on the computational performance of the tools in the evaluation process. In Light's study(Light et al., 2014), the researchers evaluated the visualization tool by checking its performance in terms of scalability. In studies that focus more on the interaction side, researchers would use metrics such as the recall accuracy (Saket, Scheidegger, Kobourov, & B¨orner, 2015) and the inference accuracy (Shen & Ma, 2008) to validate the usefulness of the prototypes.

In general, there is no universally best approach to carry out the prototyping and evaluation processes in studying data visualization. A combination of different methods that fit to the research questions would yield a better result.

Chapter 4

Phase I: Feature Engineering and Feature Selection

In this chapter, I introduce the first phase of studies I conducted to classify online users with disabilities. In the first phase, I build a binary classifier using features from three aspects of online users to explore which of them are most useful in the classification task.

To my best knowledge, no existing work has been done to classify online users with disabilities. In order to understand the problem, I drew inspirations from existing literature regarding people with disabilities and generated a set of candidate features (Table 4.1) that would be potential in classifying users with disabilities.

The features can be put into three categories. First, psychological traits, which include self-focus, cognitive process, etc. Second, community-based features, which include density, size, etc. Third, personal interests. These features have either be found different in an offline setting or been used to classify online users' other attributes such as gender, age, etc. I designed and carried out the phase one study to explore how useful these features are in classifying online users with disabilities.

Feature Name	Description	Reference
Level of self-focus	How much people focus on themselves	(De Choudhury et al., 2016)
Psychological traits	The psychological differences	(Schwartz et al., 2013)
Community sizes	The number of people in the same communities	(Brady, Zhong, Morris, & Bigham, 2013)
Community density	The density of the communities	(Brady et al., 2013)

Similar people	Number of similar people in the same community	(Wu & Adamic, 2014)

Table 4.1: Candidate features.

4.1 Data Collection and Annotation

I obtained data from Reddit[1]. Reddit is an online forum where users, also known as "redditors", can submit textual posts or other content (links, media, etc). The content is organized by topical sub-forums, which are known as "subreddits". Users can subscribe to subreddits and reply to or vote on posts and comments. One important characteristic of Reddit is that the website does not enforce the real name rule. Hence, users are anonymous, and one user could have multiple usernames. In this thesis, I treat each unique username on Reddit as a user. And use the words "user" and "username" interchangeably.

I used the official Reddit API[2] to collect posts, comments, and corresponding metadata in five steps approved by our institution's IRB. First, I collected all users who either posted or commented in two amputee-relevant subreddits (/r/amputee and /r/prosthetics) by collecting all posts and comments from the two subreddits and generating a list of usernames. Second, I manually screened and generated classes for each of the users in the list. Third, I collected all the posts (with all comments included) and comments authored by each user in the list. Fourth, I collected 1) all the comments that the users replied to, 2) all the comments that have been sent to the comments authored by the users, 3) all the posts that the users made comments to, and 4) all the comments made to the posts authored by the users. Finally, this data was used to

[1] https://www.reddit.com/
[2] https://praw.readthedocs.io/en/stable/

construct a homogeneous network, based on which I generated community-based features.

It is worth noting that, although I collected usernames from two specific subreddits, I collected the comments and posts made by each user in the dataset anywhere on Reddit.

I had help from my research lab to manually created labels ("representative" or "unrepresentative") for all the users that I collected. Two accessibility researchers labeled the users independently. I used the posts and comments submitted by each of the users to infer their class. During the process, I looked specifically for the words, sentences, or photos (Table 4.2) that users used to explicitly identify themselves as amputees.

Amputees (Representative Users)
"above elbow amputee here" "I'm a recent AKA (above knee amputee) and this was my first liner" "Really sucks being an amputee because of this" a self-taken photo of a missing leg with caption: "waking up an amputee"
Non-amputees (Unrepresentative Users)
"I am an electrical engineer working on robotic hand prosthetics in my free time" "I'm currently a 3rd year P&O student in Bundoora" "My wife is going to have a below-the-knee amputation tomorrow"

Table 4.2: Examples of self-disclosing texts in posts/comments.

For users who did not explicitly identify themselves in their submissions, the researchers made their judgment by reading through the content and looking for qualitative grounding. At last, I combined the judgments using the workflow depicted in

Figure 4.1. All the users that had inconsistent labels were excluded and their data was removed from the dataset.

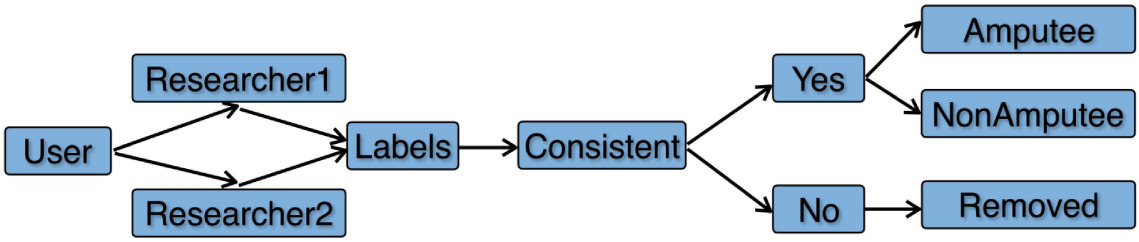

Figure 4.1: The workflow of generating class labels for online users.

Originally, I collected 752 users from the two subreddits. I removed 94 users that either have less than five submissions overall or only have submissions in the two subreddits, since such accounts could be temporarily created by users who just seek information. Based on the two researchers' labeling result (Cohen's $\kappa = 0.93$), I had 619 remaining users with 221 amputees and 398 non-amputees. These users authored a total of 40,519 posts with 242,490 comments on Reddit from June 2008 to May 2016.

4.2 Methods

To characterize the differences between representative users and unrepresentative users, I utilized measures that fall into three categories: linguistic behavior, online interactions, and community characteristics.

4.2.1 Linguistic Behavior

First, I explored each user's linguistic usage by computing the proportions of different categories of words. I use the LIWC[3] dictionary as the pre-defined categories to gauge all textual content submitted by each individual user.

[3] www.liwc.wpengine.com

Besides linguistic composition, I am also interested in the content of users' online discussions. I use the online latent Dirichlet allocation (LDA) (Hoffman et al., 2010) to derive the topic distributions for each post in the dataset. Note that I carry out the topic modeling based on posts, as I assume that each post represents a self-contained discussion session on Reddit. I included all the comments made in each post as a document. By doing this, I could extract the topics for every discussion even if the original post is not textual (e.g., a link or a photo).

Each document is tokenized, stemmed, and stopwords-removed using the Natural Language Toolkit[4] before the topic modeling process, which is carried out using the Gensim[5] package with default hyperparameters. For the purpose of interpretation, I kept only nouns in the corpus. I removed the most frequent (nouns appearing in more than 85% of the posts) and most infrequent nouns (nouns appearing less than five times overall). I first used perplexity of a held-out set of posts (20% of the corpus) to infer the range of the numbers of topics. Then I used human judgement to determine the number of topics by checking the coherence of the top ten terms of each topic. At last, I extracted 60 topics from the corpus.

4.2.2 Building a Social Network Graph

One challenge to engineer online interaction and community-based measurements on Reddit is that there is no explicit following mechanism. Thus, I need to find an alternative approach to construct a social network graph. In this study, I first build a heterogeneous graph g based on the interactions among users. In the graph g, I have three types of nodes: $u_i \in U$ represents users, $p_i \in P$ represents posts, and $c_i \in C$

[4] http://www.nltk.org/
[5] https://radimrehurek.com/gensim/

represents comments. Since on Reddit, a user i can reply to another user j by replying to either a post or a comment authored by user j, I can build edges based on the interactions by extracting the relations in Table 4.3.

	Notation	Description	Abbreviation
Node	u	Users	
	p	Posts	
	c	Comments	
Edges	$u_i \rightarrow p_k \leftarrow u_j$	user i replied to post k authored by user j	$u_i \overset{p}{\longrightarrow} u_j$
	$u_i \rightarrow c_k \leftarrow u_j$	user i replied to comment k authored by user j	$u_i \overset{c}{\longrightarrow} u_j$

Table 4.3: Interaction types in the heterogeneous graph g.

Based on these interactions, I have directed edges $r_{u_i \overset{p}{\rightarrow} u_j} \in A$ that each represent direct replies from user i to user j's posts. Also, I have $r_{u_i \overset{c}{\rightarrow} u_j} \in A$ that each represent direct replies from user i to user j's comments. Note that each of the edges is weighted by the frequencies of replies. For the edges in the graph g, I aggregate all the edges by ignoring their types. I define the transition probability from one user to another as:

$$P(u_i \longrightarrow u_j) = \frac{r_{u_i \overset{p\&c}{\rightarrow} u_j}}{\sum_{k=1}^{n} r_{u_i \overset{p\&c}{\rightarrow} u_k}}$$

In the equation, $r_{u_i \overset{p\&c}{\rightarrow} u_j}$ denotes the total number of replies from user i to user j (via both posts and comments). $\sum_{k=1}^{n} r_{u_i \overset{p\&c}{\rightarrow} u_k}$ denotes the total number of replies user i authored to the other users in the graph g. With the aggregation of edges, I created a homogeneous graph $G(V, A)$. Each $u_i \in U$ represents a user and each $a_i \in A$ represents a

directed edge with weights between two users. I will refer to this graph as graph G in our extraction of interaction and community-based measures.

4.2.3 Interaction Measures

To explore the online interaction of the two groups of users, I extract the following features: 1) the number of comments authored by each user, 2) number of posts authored by each user, 3) indegree, and 4) outdegree. I avoid using platform-specific measurements (e.g., karma scores). The indegree and the outdegree (Compeau, Pevzner, & Tesler, 2011) measurements are calculated based on the graph G.

4.2.4 Community-based Measures

Although there are predefined communities known as "subreddits" on Reddit, I believe the subreddits are too broad and too general to reveal the relationships among the users. Hence, in this study, I explore online communities by finding the groups of users that have frequent online interactions with each other. In other words, I identify communities by finding the denser parts in the graph G. To do that, I use the random walk algorithm (Pons & Latapy, 2006) on the homogeneous graph G to identify denser subgraphs, which I will treat as communities and extract the corresponding features. For each user u_i, I extract the following measurements from the community C_{u_i} that the user belongs to:

The size of a community: This feature is defined as the number of users in a community.

The density of a community: This feature is defined as

$$density(C_{u_i}) = \frac{E_{exist}}{E_{possible}}$$

E_{exist} represents the existing edges among the nodes in the community while $E_{possible}$ represents the number of all the possible edges among the nodes in the community.

The ratio of the same type of users in a community: This feature is defined as

$$ratio(C_{u_i}) = \frac{\sum u_{same}}{\sum u_{all}}$$

$\sum u_{same}$ represents the sum of the same type of users as user u_i. For example, for a representative user u_i, I detect the community C_{u_i} and calculate how many users in the community have the same label. And $\sum u_{all}$ represents the size of the community.

Personal prestige score: Since I am interested in the status of each user in his/her community, I use the PageRank algorithm (Langville & Meyer, 2011) to calculate their prestige scores.

4.2.5 Classification

Besides characterizing the differences, I also want to explore to what extent the measurements could be used to classify the two groups of users as well as which measurements are the most effective ones.

Feature selection: Before the classification task, I apply a heuristic correction (Sandri & Zuccolotto,2012) for 100 iterations to evaluate and screen the features based on the Gini index (Strobl, Boulesteix, Zeileis, & Hothorn, 2007) before the classification.

Model construction and validation: In the model construction process, I use the selected variables as predictors and the class of each user (either representative or unrepresentative) as the dependent variable to train classifiers.

To train the classifiers, I tried two popular supervised learning methods: a parametric method (the logistic regression with lasso (Tibshirani, 1996)) and a non-

parametric method (the random forest (Breiman,2001)). I compare and report the performances of the two approaches.

4.3 Findings

4.3.1 Linguistic Behavior

I applied LIWC on all the text generated by each of the users. The program counts the words and calculates the ratios of each category. Significantly different categories are shown in Table 4.4.

LIWC Dimension	**Abbrev.**	**Rep. (%)**	**Unrep. (%)**	**SIG.**
Analytical Thinking	Analytic	50.74	55.45	
Clout	Clout	45.51	53.39	***
Hearing	hear	0.58	0.58	
Authentic	Authentic	53.83	43.35	***
Emotional Tone	Tone	57.09	57.89	
Feeling	feel	0.80	0.63	***
Words longer than 6 letters	Sixltr	14.83	16.43	***
Past focus	focuspast	3.69	3.31	***
Total pronouns	pronoun	15.63	14.73	***
Personal pronouns	ppron	9.91	9.04	***
1st pers singular	i	5.86	4.52	***
Common adverbs	adverb	5.56	5.39	*
Work	work	1.76	2.34	***
Affect Words	affect	6.00	5.64	***
Negative emotion	negemo	2.08	1.85	*

Sadness	sad	0.41	0.33	***
Biological Processes	bio	2.91	2.35	***
Body	body	1.27	0.87	***
Health/illness	health	0.99	0.73	***
Discrepancies	discrep	1.79	1.95	***
Time	time	4.55	4.16	***

*** adjusted $p < .001$ ** adjusted $p < .01$ * adjusted $p < .05$

Table 4.4: Differences on LIWC categories between representative and unrepresentative users. Statistical significance is based on Wilcoxon Rank-Sum test with Holm-Bonferroni adjustment.

From the results, I found that there is a different attentional focus between representative users and unrepresentative users. Representative users use more self-references (1st pers singular, $w = 60612.5$) than unrepresentative users, which is consistent with previous research findings that people who are experiencing emotional or physical pain tend to focus more on themselves (Rude et al., 2004). Also, representative users are more past-focused ($w = 54006.5$), which is typical when people discuss events they have previously disclosed about (Pasupathi, 2007). Second, I found unrepresentative users to be potentially more extroverted. Overall, unrepresentative users show less negative emotions ($w = 51306$) and sadness ($w = 54369.5$), which were found to be correlated to higher scores in extroversion (one of the big-five personality traits) (Pennebaker& King, 1999; Mehl, Gosling, & Pennebaker, 2006).

Besides potential psychological differences, there were difference in content-related dimensions. Representative users discuss more health-related content (body, $w = 59088.5$; health/illness,$w = 55588$) while unrepresentative users' discussion is more work oriented (work, $w = 29769.5$).

4.3.2 Content Analyses

To further explore the content differences, I constructed an LDA model and calculate the topic proportions for each of the posts in the corpus.

Based on the topics extracted using the LDA model, two researchers created an affinity diagram based on a corpus consists of the top 10 posts with the largest topic proportion from each topic. The main purpose is to extract overarching themes. As shown in Table 4.5, I derived ten themes, which are health, job and finance, URLs, entertainment, politics, life, home, travel, religion, and electronics.

As I compare the average topic distribution (Table 4.5), I found that representative users post more physical disability-related content (Topic 11). Unrepresentative users focus more on the treatment aspect of disabilities (Topic 9). As I read through the corresponding posts, I found such content was largely contributed by medical practitioners. Besides these, I also found unrepresentative users posted more work and life-related content (Topic 26, 46).

Theme	**Topic IDs**	**Top 5 Terms**	**Rep. (%)**	**Unrep (%)**	**SIG.**
Health	3	food, eat, day, calorie, week	0.79	0.90	
	11	leg, prosthetic, foot, year, pain	3.82	2.72	*
	20	hand, arm, thing, work, finger	3.50	2.62	
	53	body, weight, loss, blood, time	1.35	1.14	
	9	drug, company, money, people, cost	1.86	2.64	*
Electronics	21	amp, circuit, breaker, row, gpu	0.26	0.33	

	29	phone, test, minute, plan, work	1.13	1.02	
	31	data, use, program, camera, app	0.75	0.79	
	50	Use, power, work, design, model	1.91	2.48	
Job & Finance	19	people, job, law, work, state	1.12	1.48	
	26	use, work, problem, file, number	2.08	2.93	*
	37	tax, price, trade, sale, income	0.54	0.77	
	49	work, time, job, day, thing	3.72	5.04	
	54	money, account, card, bank, credit	0.92	1.19	
	57	use, work, thing, hand, time	2.36	2.24	
Religion	51	people, god, belief, person, thing	1.88	1.55	
URLs	12	com, http, imgur, jpg, thank	2.92	2.57	
	14	com, http, www, music, video	3.76	3.41	
	16	http, com, razor, www, soap	0.33	0.36	
	17	dfg, com, http, look, car	1.39	1.33	
	36	r, com, reddit, http, comment	3.63	3.94	
Travel	35	flight, plane, fly, air, pilot	0.53	0.69	
	1	stream, flood, resolute, podcast, fire	0.97	1.23	
Enterta-inment	2	game, play, player, time, pc	2.56	2.26	
	4	sword, character, armor, tank, box	1.15	1.03	
	7	race, x, download, dragon, species	0.34	0.46	
	8	movie, guy, time, girl, thing	4.76	3.67	
	28	spoiler, review, soap, japan, tag	0.65	0.61	
	43	hero, card, use, draft, mission	1.29	1.31	

	47	site, eu, teel, chip, island	0.56	0.61	
	52	use, damage, attach, item, level	0.89	0.85	
Home	25	dog, year, cat, month, home	1.62	1.21	
	27	date, moumouren, origin, amount, thread	0.49	0.59	
	32	water, tank, seed, use, filter	0.76	0.67	
	33	color, hair, look, paint, use	1.43	1.07	
	39	car, time, park, drive, road	2.50	2.73	
	56	bag, beer, drink, tea, sock	0.81	0.83	
	59	cook, egg, use, recipe, chicken	1.11	1.13	
	60	phone, wife, glass, carry, mon	0.86	1.02	
Politics	18	world, war, crime, power, state	0.97	1.23	
	24	people, vote, news, polite, support	1.27	1.42	
	30	shoot, gun, time, order, rifle	1.24	1.16	
	41	ban, gun, prohibit, waster, femin	0.39	0.40	
	42	name, guy, http, man, com	1.93	1.81	
	48	gun, use, word, weapon, fire	0.79	0.81	
Life	5	server, bowl, use, map, host	0.38	0.49	
	6	area, city, bird, wire, control	1.43	1.41	
	10	people, thing, way, talk, time	3.39	2.92	
	13	day, year, time, night, week	2.61	2.34	
	15	women, men, woman, male, gender	0.63	0.68	
	22	friend, family, brother, roll, parent	1.36	1.42	
	23	book, dream, horse, baby, life	0.86	0.75	

	44	photo, facebook, day, time, picture	1.22	1.41	
	46	school, year, college, class, student	0.81	1.39	***

*** adjusted p < .001 ** adjusted p < .01 * adjusted p < .05

Table 4.5: Themes in posts and corresponding topics with top 5 terms. Statistical significance is based on Wilcoxon Rank-Sum test with Holm-Bonferroni adjustment.

Overall, I found that posts submitted from the two groups covered a large range of themes. However, representative users' discussions are more health-oriented while unrepresentative users' are more job and finance oriented. The finding is consistent with the LIWC results.

4.3.3 Online Interaction

Based on the comparison of the online interaction measurements, I found that representative users have higher indegree, and they generate fewer posts in total. The results are shown in Table 4.6 (left half).

Interaction	**Rep.**	**Unrep.**	**SIG.**	**Community**	**Rep.**	**Unrep.**	**SIG.**
# Comments	357.69	408.60		Size	647.23	663.42	
# Posts	41.27	78.50		Rep User Ratio	0.02	0.01	***
Indegree	373.29	305.23	***	Unrep User Ratio	0.01	0.03	***
Outdegree	262.12	267.50	***	Density	0.03	0.04	
				Prestige Scores	0.36	0.37	

*** adjusted p < .001 ** adjusted p < .01 * adjusted p < .05

Table 4.6: Online interaction and community features test using Wilcoxon Rank-Sum test with Holm-Bonferroni adjustment.

4.3.4 Community Characteristics

I constructed a homogeneous network with 234,386 nodes and 329,592 weighted directed edges. Note that I include not only the users from the two subreddits but all the

users who authored replies on Reddit. I applied random walk to identify online communities. For each of the users, I extract five features based on the community that user belongs to. The results of the comparisons are shown in Table 4.6 (right half).

I did not find significant differences between the two groups of users in terms of community size, density, or prestige score. However, I found that representative users have more representative users in their communities than unrepresentative users. Also, we did the same comparison of the ratio of unrepresentative users in both groups and found the same pattern. This finding supports our hypothesis of homophily.

4.4 Classification Results

I used all the features extracted from linguistic behavior, online interaction, and community analysis to classify the users. I first carried out a feature selection process to reduce the total number of candidate predictors. Then I applied two supervised learning algorithms, logistic regression with LASSO (LR) and random forest (RF) to train classifiers. The performances of the two methods and most useful predictors are reported in the section.

Feature	Category	Importance	Feature	Category	Importance
Rep user ratio	Community	100.00	ppron	LIWC	0.64
Unrep user ratio	Community	34.63	discrep	LIWC	4.52
i	LIWC	6.97	Topic 8	LDA	0.24
body	LIWC	5.56	pronoun	LIWC	1.88
Authentic	LIWC	4.72	sad	LIWC	0.93
work	LIWC	8.06	time	LIWC	1.76
Clout	LIWC	5.98	focuspast	LIWC	0.70

Topic 11	LIWC	0.99	Topic 26	LDA	1.05
Sixltr	LDA	5.15	negemo	LIWC	1.70
bio	LIWC	2.31	Topic 20	LDA	1.90
health	LIWC	1.73	feel	LIWC	0.00
Topic 46	LIWC	2.11	affect	LIWC	0.71
Topic 49	LDA	1.19	#posts	Interaction	1.02
*** adjusted p < .001 ** adjusted p < .01 * adjusted p < .05					

Table 4.7: Selected features and relative importance in full model.

4.4.1 Feature Re-calculation

I re-calculate two community features–representative user ratio and unrepresentative user ratio–using the label information only in the training dataset for all instances. And the updated values are used in the classification task.

4.4.2 Feature Selection

For each user, I derived a total number of 161 predictors, which include 92 LIWC dimensions, 60 topic proportions, 4 online interaction features, and 5 online community features. To reduce the dimensionality of the dataset, I adopted the heuristic Gini index correction approach to screen the features. Specifically, I went through 100 iterations of the correction process. In each iteration, I set the number of trees to grow to500. I choose the cutoff point to be a common value of 0.5, which gives us 26 remaining features (Table 4.7) for the classification task.

4.4.3 Training and Testing

I randomly sample 30% of the data as the testing set (66 representative user and 119 unrepresentative users).The remaining 70% was used as the training dataset (155 representative users and 279 unrepresentative users).In order to improve the performance

of the classifiers, I apply under-sampling techniques on the training dataset to make the number of instances in each class balanced. For the testing processing, I keep the original proportion of the two classes. To tune the models, I use repeated 10-fold cross-validation during the training process and selected the final model based on accuracies. The testing dataset was applied at last to validate the models.

Our result shows the RF model performs better in both sensitivity (91% vs. 64%) and specificity (86%vs. 70%) than the LR model (Figure 4.2). Thus, I chose it to be our final model for interpretation. To rank the features in the model, I use variable importance based on the Gini index (Calle & Urrea, 2011) to quantify their contributions. The relative importance (ranging from 0 to 100) of each predictor is shown in Table 4.7. Furthermore, I build separate models using subsets of features to explore their performance (Table 4.8). I can see that community features are more useful in identifying representative users than linguistic features. The fact shows strong evidence of the Homophily phenomenon.

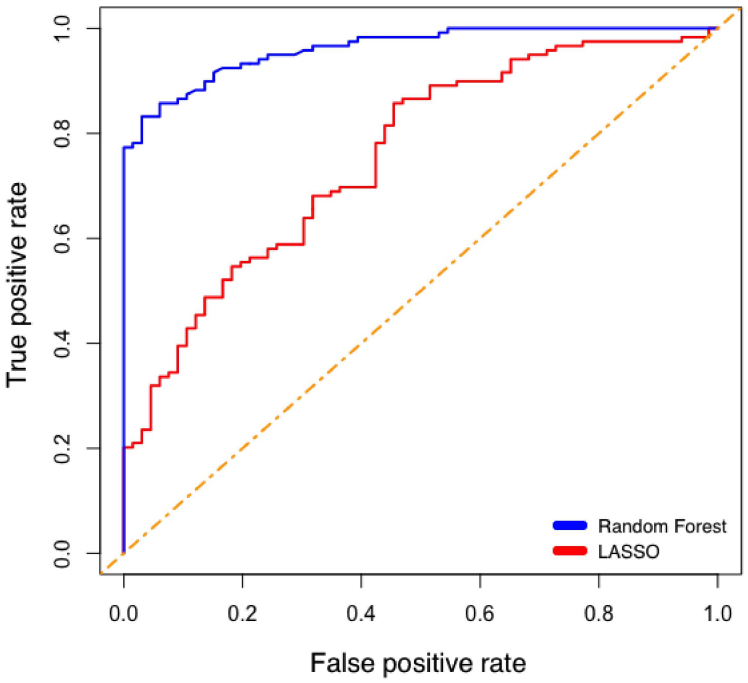

Figure 4.2: ROC curves comparison between random forest and LASSO models.

4.5 Phase One Conclusion

The main purpose of the phase one study is more about exploring useful input signals than the performance of the classification task. I have several limitations in the designing of the classification solution. First, I compiled a relatively small dataset with 619 online users. Second, the sampling of the data is biased that I only collected users with physical disabilities from two related subreddits. Third, the classification task was binary, which will not support multiple classes classification. Fourth, the overall computation cost is expensive, since I applied different algorithms in the feature engineering process to obtain the features. From the findings, I also gain inspirations what a better classifier would look like. First, the manual annotation process was expensive and formidable. It would be ideal to have classifier that does not require a lot of train data to perform well. Second, the overall computation cost should be improved.

Third, the model should support multi-class classification. With the goals in mind, I designed and carried out a phase two study to devise a better classification model.

Chapter 5

Phase II: A Co-training Model with Label Propagation on a Bipartite Graph

In this chapter, I propose a semi-supervised co-training model to identify disabled users on Reddit.com. The proposed model is devised based on the assumption of homophily, which presumes that online users with a same disability are closely tied to each other via disability-related online posts in their online social networks. The model uses a variational label propagation algorithm to capture the social network information, and an auxiliary classifier to capture the textual information in online posts. I carried out experiments based on a larger and newly compiled dataset collected from Reddit.com and presented the results, which showed that the new approach can classify users with disabilities with greater performance than baseline methods that include both text-based and graph-based classification models.

5.1 Assumption of the Co-training Model

Our fundamental assumption is that sharing the same disability is a potential dimension of homophily. Hence, users on social media websites are more likely to interact with each other in posts that are discussions of disabilities if they share a same disability. Although non-representative users may also participate in these discussions, their participation are less frequent and persistent according to the homophily principle. If I can create a graph that captures differences of social ties among online users, I could use the label propagation algorithm to find users who are closely related to other online users with the same disability.

As aforementioned in the related work section, the label propagation algorithm is an effective machine learning method. First, it is a semi-supervised learning method that

allow data to learn from unlabeled data as well as labeled data. This could reduce the cost required to annotate a sufficient dataset. Second, the computation cost is inexpensive since it is based on matrix multiplications. These two traits made label propagation an ideal candidate solution for the task at hand. Nevertheless, I observed two problems that will make direct adoption of the method not ideal. First, the homophily principle suggests that online users' interactions are based on multiple dimensions. Second, the real-world social network is always sparse. And I will discuss these two problems in details.

To begin with, I want to formalize the intuition of homophily, let's define an undirected graph $G(V, E)$ where $v_i \in V$ represents an online user and $e_{ij} \in E$ represents an edge that connects v_i and v_j on a social media website. W is a weight function that returns the edge weight $w_{ij} = W(e_{ij})$, which is a quantitative measurement of the observed frequency of online interactions between these two online users (e.g., their exchange of replies). It is natural to assume that $W(e_{ab}) > W(e_{ac})$ if $\phi(v_a) = \phi(v_b)$ and $\phi(v_a) \neq \phi(v_b)$, where $\phi(v_i)$ is a function that gives the class label of v_i.

However, according to the homophily principle, an individual's social network is a complex that is based on different levels on different dimensions of homophily (McPherson et al., 2001). Hence, the assumption that $W(e_{ab}) > W(e_{ac})$ does not always hold for $\phi(v_a) = \phi(v_b)$ and $\phi(v_a) \neq \phi(v_c)$. For example, as I use the observed frequency of online interactions among users to predict online users with disabilities. I could observe that $W(e_{ab}) < W(e_{ac})$ in the scenario that v_a and v_c has another strong homophily on a dimension such as a mutual hobby (e.g., they are both interested in cars).

Hence, a method to let label information propagate only via desired relationship is necessary to reduce classifications errors.

Besides the problem of multiple dimensions of homophily, it is also challenging to choose a set of labeled vertices in a graph to propagate with while using a real-world social network with label propagation. A graph that represents a real social network is usually sparse and disconnected. With the initial nodes chosen randomly, it typically yields bad results since label information cannot propagate through disconnected regions in a graph. The classification results will have low recalls. In order to improve the performance, modifications are necessary for the label propagation algorithm. To measure the sparsity of a graph, I used density to check the sparsity of a graph. Density is the ratio between the number of observed edges and number of all possible edges in a graph. The number ranges from 0 to 1 where 1 means a fully connected graph. The density of the social network in study 1 is 0.011, which means the graph is very sparse in the dataset.

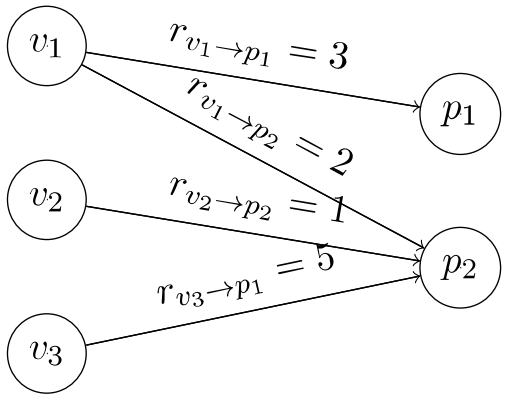

Figure 5.1: An example of a bipartite graph that contains user nodes V and post nodes P.

5.2 A Bipartite Graph Representation

To address the first problem that online users' interactions are based on the combination of different homophily dimensions, I proposed a new way to represent online users. Instead of using a homogeneous graph, which was introduced in phase I, I proposed to use a bipartite graph representation. In this graph, there are two types of nodes. One representing online users while the other representing online posts. User nodes are connected via post nodes. I made this assumption that each online post has a clear topic and represents one type of homophily dimension. By incorporating post nodes, I separate the observation of online interactions between a pair of user nodes into multiple paths connected by post nodes. This would allow label information to propagate via different posts nodes and then back to user nodes. In order to make the label propagation algorithm work on a bipartite graph, I proposed a modified version.

To introduce the variation of the label propagation algorithm, I formally define a bipartite graph $G_B(V, P, A)$, which represents the social networks of online users (Figure 5.1). In G_B, V are user nodes, and P are post nodes. $a_{ij} \in A$ are the directed edges that connect the two types of vertices. Since user nodes V are connected via post nodes P in this graph, an edge a_{ij} always points from a v_i to a p_j.

Given the graph G_B, I define two functions for calculating transition probabilities for edges:

$$W_{v \to p}(a_{ij}) = \frac{r_{v_i \to p_j}}{\sum_{k=1}^{|P|} r_{v_i \to p_k}}$$

The function $W_{v \to p}$ returns the transition probability of an edge $a_{ij} (i \neq j)$ that points from v_i to p_j. $r_{v_i \to p_j}$ denotes the count of comments authored by v_i in post p_j. For

example, if v_i left three comments in post p_j, then $r_{v_i \to p_j} = 3$. $|P|$ denotes the cardinality of P.

$$W_{p \to v}(a_{ij}) = \frac{r_{v_i \to p_j}}{\sum_{k=1}^{|V|} r_{v_k \to p_j}}$$

The function $W_{p \to v}$ returns the transition probability of the same edge a_{ij} from an opposite direction. $|V|$ denotes the cardinality of V.

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

$$T_{ij} = \begin{cases} W_{v \to p}(a_{i(j-|V|)}) & \text{if } i \leq |V| \text{ and } j > |V| \\ W_{p \to v}(a_{(j-|V|)i}) & \text{if } i > |V| \text{ and } j \leq |V| \\ 0 & \text{otherwise} \end{cases} \quad (5.1)$$

There are two things worth noting regarding the bipartite graph. First, rows of the adjacency matrix T are normalized, which is important in proving that the label propagation algorithm would converge in the next section. Second, in the bipartite graph, I made user nodes connected via post nodes. The purpose is to separate connections among users via different posts. Given our previous assumption of homophily, user nodes with the same labels are more likely to be connected to each other via certain post nodes, which are typically discussions on disability related topics. And I want the label information to propagate via paths consist of these post nodes.

5.3 Label Propagation on a Bipartite Graph

With the bipartite graph G_B established, I introduce the label propagation algorithm in this section. In this variation of the algorithm, two types of vertices, V and P in G_B, both have labeled and unlabeled sets. And the possible labels L for the two types of nodes are the same.

Given a bipartite graph $G_B(V, P, A)$, I define an adjacency matrix $T \in R^{n \times n}$ and a label matrix $C \in R^{n \times k}$, in which $n = |V| + |P|$ and $k = |L|$ that $L = l_1, l_2, l_3, \ldots l_k$ is the set of all possible labels. $C_{ij} = 1$ if v_i or p_i has label l_j and $C_{ij} = 0$ otherwise. Intuitively speaking, for $v_i \in V$, the label vector C_i represents the probabilities of having corresponding disabilities for v_i. For $p_i \in P$, the label vector C_i represents the probabilities of users who participated in this post have that disabilities. For example, if a post has high probability of the class "amputee", then users who participated in the post are more likely to be amputees.

The set V and P are both separated into a labeled and an unlabeled set, which are denoted as V^l, V^u and P^l, P^u. And $V^l \cup V^u = V$, $P^l \cup P^u = P$. The goal is to learn the labels of V_u in G_B. Since all $v_i \in V$ are connected via $p_j \in P$, I can learn them simultaneously using a modified label propagation algorithm. The label information in C propagates based on a transition matrix as defined in equation 5.2:

$$\begin{bmatrix} C_{V^l} \\ C_{V^u} \\ C_{P^l} \\ C_{P^u} \end{bmatrix} := \begin{bmatrix} T_{V^lV^l} & T_{V^lV^u} & T_{V^lP^l} & T_{V^lP^u} \\ T_{V^uV^l} & T_{V^uV^u} & T_{V^uP^l} & T_{V^uP^u} \\ T_{P^lV^l} & T_{P^lV^u} & T_{P^lP^l} & T_{P^lP^l} \\ T_{P^uV^l} & T_{P^uV^u} & T_{P^uP^l} & T_{P^uP^u} \end{bmatrix} \cdot \begin{bmatrix} C_{V^l} \\ C_{V^u} \\ C_{P^l} \\ C_{P^u} \end{bmatrix} \quad (5.2)$$

In equation 5.2, T is the normalized adjacency matrix of G_B derived based on equation 5.1. In each iteration, the class information C is propagated and updated based on the adjacency matrix of the graph. The subscriptions denote parts of the matrix (e.g., $T_{V^lV^u}$ denotes transition probability matrix between V^l and V^u). The sub-matrices $T_{V^lV^l}$, $T_{V^lV^u}$, $T_{V^uV^l}$, $T_{V^uV^u}$, $T_{P^lP^l}$, $T_{P^lP^u}$, $T_{P^uP^l}$, and $T_{P^uP^u}$ are matrices of 0s due to the fact that the same type of nodes don't have edges among themselves in the bipartite graph G_B.

Hence, I am only interested in the label information that propagates across the two types of vertices in each iteration:

$$C_{V^u} \coloneqq T_{V^u P^l} C_{P^l} + T_{V^u P^u} C_{P^u}$$

$$C_{P^u} \coloneqq T_{P^u V^l} C_{V^l} + T_{P^u V^u} C_{V^u}$$

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

$$C_{V^u}^{(n)} = \sum_{i=0}^{n-1} (T_{V^u P^u} T_{P^u V^u})^i \left(T_{V^u P^l} C_{P^l} + T_{V^u P^u} T_{P^u V^l} C_{V^l}\right) + (T_{V^u P^u} T_{P^u V^u})^n C_{V^u}^{(0)} \quad (5.3)$$

$$C_{P^u}^{(n)} = \sum_{i=0}^{n-1} (T_{P^u V^u} T_{V^u P^u})^i \left(T_{P^u V^l} C_{V^l} + T_{P^u V^u} T_{V^u P^l} C_{P^l}\right) + (T_{P^u V^u} T_{V^u P^u})^n C_{P^u}^{(0)} \quad (5.4)$$

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

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

Based on this constraint, the following can be proven:

$$\sum_{j}^{|V^u|} B^n[i,j] = \sum_{j}^{|V^u|} (B^{n-1}B)[i,j]$$

$$= \sum_{j}^{|V^u|} \sum_{k}^{|V^u|} B^{n-1}[i,k] B[k,j]$$

$$= \sum_{k}^{|V^u|} B^{n-1}[i,k] \sum_{j}^{|V^u|} B[k,j]$$

$$\leq \sum_{k}^{|V^u|} B^{n-1}[i,k]\gamma \leq \gamma^n, \forall i = 1,2,3,\dots,|V^u|$$

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

$$\lim_{n\to\infty}(T_{V^uP^u}T_{P^uV^u})^n C_{V^u}^{(0)} = 0, \lim_{n\to\infty}(T_{P^uV^u}T_{V^uP^u})^n C_{P^u}^{(0)} = 0 \quad (5.5)$$

It is clear, by plugging equations 5.5 back into equations 5.3 and 5.4, the results of label propagation do not depend on the initial value of C_{V^u} and C_{P^u}. The algorithm will converge eventually based on the adjacency matrix T and label matrix C as long as they are row normalized.

I refer to this variation as label propagation on a bipartite graph (LPBG). The details are summarized in Algorithm 1 (refer to Appendix A for the implementation in Python3). The algorithm returns the probability matrix C at termination, which can be used for class assignments.

With the new algorithm LPBG, I can apply label propagation on a bipartite graph representation. But the second problem still remains, which is the fact that online social network, whether represented as a bipartite graph or not, is still spare. And the label propagation algorithm will not propagate label information to a disconnected area in such a graph. To mitigate this problem, I propose a co-training model in the next section.

Algorithm 1: Label propagation on a bipartite graph (LPBG)

1 //Inputs:

2 Adjacency matrix T

3 Labeled sets $V^l$and P^l

4 Label matrices $C_{V^l}^{(0)}$and $C_{P^l}^{(0)}$

5 Convergence condition ϵ

6 // Label propagation process

7 Initialize $C_{V^u}^{(0)}$ and $C_{P^u}^{(0)}$as matrices of 0s

8 **while** *True* **do**:

9 // clamp learning sets

10 $C_{V^l} \leftarrow C_{V^l}^{(0)}$

11 $C_{P^l} \leftarrow C_{P^l}^{(0)}$

12 // propagate information to unlabeled nodes

13 $C_{V^u}^{(n)} \leftarrow T_{V^u P^l} C_{P^l} + T_{V^u P^u} C_{P^u}^{(n-1)}$

14 $C_{P^u}^{(n)} \leftarrow T_{P^u V^l} C_{V^l} + T_{P^u V^u} C_{V^u}^{(n-1)}$

15 // row normalize $C_{V^u}^{(n)}$and $C_{P^u}^{n}$

16 $C_{V^u}^{(n)} = normalize\left(C_{V^u}^{(n)}\right)$

17 $C_{P^u}^{(n)} = normalize\left(C_{P^u}^{(n)}\right)$

18 // check for convergence

19 **if** $\left|\left|C_{V^u}^{(n)} - C_{V^u}^{(n-1)}\right|\right| + \left|\left|C_{P^u}^{(n)} - C_{P^u}^{(n-1)}\right|\right| < \epsilon$ **then**

20 **return** $C_{V^u}^{(n)}$, $C_{P^u}^{(n)}$

21 **else**

22 continue

23 **end**

24 **end**

5.4 Co-Training with Label Propagation

In this section, I introduce our design of co-training model that leverages LPBG and a naive Bayes classifier (NBC). The idea of co-training assumes that each data instance X in the dataset has two representations, which are denoted as $< x_1, x_2 >$. Each representation should contain sufficient information to learn the label of X and $x_1 \neq x_2$. In our design, each post node $p_i \in P$ has two representations. x_1 is the bipartite graph G_B (introduced in section 3.2) that contains the network information. x_2 is the textual content in each post.

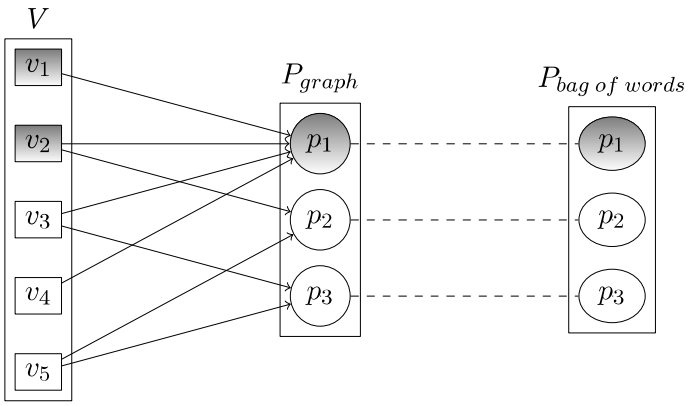

Figure 5.2: Diagram of the co-training model. Rectangles nodes represent users, circle nodes represent posts. Elliptical nodes are the textual information of posts. Shaded nodes belong to the labeled set. Un-shaded nodes belong to the unlabeled set.

The structure of the two representations in the co-training model is depicted in Figure 5.2. On the left side is the bipartite graph representation of posts. On the right side is the textual representation of posts. Each pair of representations is connected via a dashed line, which indicates that they are the same post. A set of posts will be chosen as

the labeled set, represented as shaded nodes in Figure 5.2, for the co-training model to train.

I use LPBG as the first classifier f_1 on representation x_1, P_{graph} in Figure 5.2, and a naive Bayes classifier, which is commonly applied in text classification (Lewis & Ringuette, 1994), as the second classifier f_2 on representation x_2, $P_{bag\ of\ words}$ in Figure 5.2. The model trains classifiers f_1 and f_2, using LPBG and NBC independently, on the two representations in each iteration. Each of the newly trained classifiers learns k most confident instances for each class from the unlabeled set P^u. The $|L| \times k$ newly learned instances will be removed from P^u and added into P^l for the next iteration. When sufficient label information is learned, the model uses LPBG on the bipartite graph representation to learn the final labels of online users.

As aforementioned, the co-training model trains a naive Bayes classifier to help learn the labels of online posts P based on textual information in addition to the LPBG algorithm. This design has two benefits. First, the NBC generates labeled nodes, no matter if they are connected to the labeled set, after each iteration. Second, by combining the two algorithms, the model remains a semi-supervised learning method, which helps save the effort of manual labeling.

The inputs of the co-training model consist of a bipartite graph (denoted as G_B) and a corpus of online posts (denoted as D). The bipartite graph G_B is the same one described in section 3.2. The corpus D is a set of posts. Each document $d_i \in D$ contains all the words in the post p_i. A preprocessing step is carried out on the corpus D that include tokenizing, stemming, and stop-words removing.

Algorithm 2: Co-Training Model with LPBG

//Inputs:

A bipartite graph $G_B(V, P, A)$, Corpus D

// Initialization

Calculate degrees for user vertices V in G_B

Stratified random sampling on V based on degrees to generate V^l and $C_{V^l}^{(0)}$

Collect post nodes connected to $v_i \in V^l$in $G_B$to generate P^l

Generate labels $C_{P^l}^{(0)}$from P^l

Generate subset D_{P^l} from D

Set overall number of instances to learn at each interaction K_{all}.

Calculate number of instances for each class to learn $K = \{k_1, k_2, \dots, k_{|L|}\}$ at each iteration based on the proportion of each class in V^l

// start co-training

$i = 0$

set number of interaction m

while $i < m$ **do**

  Train $f1$ using LPBG(algorithm 1)

  Train $f2$ using the naïve Bayes algorithm and D_{P^l}

  learn k_i instances with $f1$ for each class in L

  learn k_i instances with $f2$ for each class in L

  remove from P^u and add K_{all} instances into the P^l set

  add corresponding posts from D into D_{P^l} according to the newly learning instances

  $i = i + 1$

  end

Run LPBG (algorithm 1) and return C_{V^u} and generate labels based on C_{V^u}

5.4.1 Initialization Process

The initialization process aims to generate V^l and P^l as the training-set for the co-training model. The process is summarized as step 4 to 10 in algorithm 2.

The summation of counts of comments/posts that a user authored in total is used as the degree for a user node, denoted as $Degree(v_i)$, in G_B. Then a normalization of the degrees of all users is proceeded as following:

$$Fr_{v_i} = \frac{Degree(v_i)}{\sum_k^{|V|} Degree(v_k)}$$

After deriving Fr for each user node, the process uses stratified random sampling to get users from each of the six classes as the training-set V^l based on Fr. Intuitively speaking, online users who participated in more posts would have high chances of being sampled into the training-set. Their label information would help us generate a sufficient amount of labeled posts for the naive Bayes classifier.

Based on the chosen set of V_l, the process collects all $p_i \in P$ in G_B that are connected to $v_i \in V^l$. The proportion of different types of users, denoted as S_{p^l}, that participated in each online post is calculated as following:

$$S_{P^l}[i,j] = \frac{\sum_{k=1}^{|V_{p_i}|} I\{\emptyset(v_k)=l_j\}}{|V_{p_i}|} \quad (5.6)$$

$$C_{P^l}^{(0)}[i,j] = \begin{cases} 1 \text{ if } \max\left(S_{P^l}[i]\right) = S_{p^l}[i,j] \\ 0 \qquad\qquad\qquad\quad \text{else} \end{cases} \quad (5.7)$$

I is an indicator function that returns 1 if the expression inside is true and 0 otherwise. $\emptyset$ is the function that returns labels for user nodes. V_{p_i} is the set of users that are connected to p_i. And the max function returns the maximum value in a vector.

After calculating the matrix S_{P^l}, the process generates label matrix $C_{P^l}^{(0)}$ for P^l based on equation 5.7. An under-sampling process is carried out to balance the instances with different labels in P^l to improve the performance of the NB classifier. Then I extract $d_i \in D$ that match $p_i \in P^l$ to create a subset corpus D_{P^l}, which contains all the words in posts of the training-set. This corpus is used as the training data from the naïve Bayes classifier.

It is worth noting that there is a third group of users in G_B who participated in at least one post $p_i \in P$ but don't have labels. The groups of users are included when generating the graph to keep the edge weights and proportions of users in posts accurate. But these are not considered in training the NB classifier.

After this point, V^l ,V^u , P^l, P^u, $C_{V^l}^{(0)}$,$C_{P^l}^{(0)}$, and D_{P^l} are already for the co-training process. The initialization process only runs once. After it is done, all parameters are passed as references to the co-training process and will be updated accordingly.

5.4.2 Co-Training Process

The co-training process is summarized in step 11 to 22 in algorithm 2. After the initialization process, m(pre-defined) iterations of training are executed. In each iteration, LPBG and NBC are applied to train two classifiers independently. At the end of each iteration, each of the two classifiers learns k (also pre-defined) instances for each of the

classes in L. The top k most confident instances in each class are generated by choosing the top k instances with the highest probability of that class.

In training the NB classifier, each document $d_i \in D_{P^l}$ is represented as a term frequency – inverse document frequency (TF–IDF) vector. Since it is a bag-of-words based representation, order of words is ignored. At the end of each iteration, V^l, V^u, P^l, P^u, $C_{V^l}^{(0)}$, $C_{P^l}^{(0)}$ and D_{P^l} are updated and passed to the next iteration.

5.4.3 Final Labels

After m iterations, steps 23 in Algorithm 2 generate final labels for V^u. A final iteration of LPBG is carriedout based on the update inputs, which returns C_{V^u}. Then label for $v_i \in V^u$ are derived by choosing the one class with the highest probability (Equation 5.8).

$$\emptyset(V_i) = L\left[\underset{\text{j}}{argmax}\, C_{V^u}[i,j]\right] \quad (5.8)$$

To evaluate the performance of our proposed model, I collected and compiled a new dataset fromReddit1in the experiment. I collected all reddit users, who are also known as Redditors, that submitted posts/comments in at least one the following disability related subreddits: r/disability, r/amputee, r/prosthetics, r/autism, r/epilepsy, r/blind, r/deaf Blind. Based on the usernames, I collected all data generated by each of the users.

In the manual annotation process, two accessibility researchers manually annotated each online user based on user-generated text, pic, videos, flairtext (short texts that users add to their usernames in subreddits), etc. The goal is to look for any information that users self-disclosed to identify their disabilities. All the users in the final

dataset have explicitly identified themselves as either having disabilities or close to someone who has.

Finally, I annotated 3,644 online users (Cohen's \kappa=\ 0.84) that include five different types of disabilities (Table 5.1). There are 1,065 Redditors who identified themselves as someone who are close to disabled people. They are family members, caretakers, or practitioners who work in the accessibility domains and soon. This group of users are collectively known as non-representative users in the dataset.

Class	**# Users**
Autism	1168
Visually Impaired	70
Epilepsy	1093
Multiple Sclerosis	27
Amputees	221
Non-Representative	1065
Total	3644

Table 5.1: Results of manual annotation (Cohen's $\kappa = 0.84$).

Based on the annotated usernames, I collected 2,601,992 posts that these users authored/commented with all comments included from 2008 to 2019. I also collected all the usernames that participated in these posts for graph generation in the experiment.

Class	Precision	Recall	F1	#Users
Autism	0.72	1.00	0.84	1168
Visually Impaired	0.62	0.89	0.73	70
Epilepsy	0.93	0.94	0.93	1093
Multiple Sclerosis	1.00	0.86	0.92	27
Amputees	0.85	0.61	0.71	221
Non-Representative	0.92	0.55	0.69	1065
Macro Avg.	0.84	0.81	0.80	

Table 5.2: Classification metrics for each class using the co-training model and 75% data as the training-set ($K_{all} = 600$,m= 10). The overall accuracy (macro average) is 0.82.

5.5 Classification Results of the Model

The classification results of the co-training model are summarized in Table 5.2. The metrics were derived as an average of five runs of the model. At each run, 75% of the data is used as the training-set while the rest is used as the testing set to calculate the metrics. At last, I have an overall accuracy of 82% (macro average over all classes) with a precision of 84% and a recall of 81%. It is worth noting that despite the classes are unbalanced, as shown in the #Users column in Table 5.2, the new model showed good overall performance in the experiment.

5.6 Comparison with Baselines

I selected baseline models to carry out the same classification task in our experiment. Their performances are summarized in Table 5.3. Baselines 1-3 are common text classification methods, which I used to classify online user based on user-generated text. In baselines 4-6, I selected graph-based algorithms to perform the classification task.

Baselines	Precision	Recall	F1	Acc.
LDA	0.49	0.51	0.46	0.49
TF-IDF	0.51	0.66	0.53	0.66
Word2Vec	0.59	0.63	0.60	0.61
Label Prop.	0.57	0.05	0.09	0.65
Node2Vec	0.72	0.52	0.44	0.64
Network&Attribute	0.60	0.65	0.61	0.67
Co-training	0.84	0.81	0.80	0.82

Table 5.3: Comparison between the co-training model and the baselines (macro avg.). The performance is based on using 75% of the data as the training-set.

5.6.1 Baseline 1: LDA

In baseline 1, I used Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) to analyze texts generated by users. I first applied LDA to learn topics distribution for each user based on the posts their authored/commented. Then the mean topic distribution vectors are used to classify the users using a supported vector machine (SVM) with a RBF kernel. I applied the one versus all strategy while training the SVM classifiers since it is a multiclassification task. A pre-processing process with tokenizing, stemming, and stop-words removing was carried out before the LDA procedure. Most frequent term (appeared in more than85% posts) and most infrequent (appeared less than 5 times) are removed from the corpus as well.

In the LDA process, I used the perplexity of a held-out set that contains 20% of the corpus to infer the number of topics. Also, I manually checked the coherence of the top ten most frequency words of each topic in the LDA result to ensure the quality of the resulting topics. Finally, 60 topics were extracted. Each post is assigned with a vector of

topic distribution. For each user, I calculate the mean topic distribution vector of all the posts that user author/participated. Then the mean topic distribution vector is used as the input.

5.6.2 Baseline 2: TF-IDF

In baseline 2, I used term frequency-inverse document frequency (TF-IDF) vector to represent the posts. The overall procedure is the same as Baseline 1. I carried out the same pre-processing process. At last, each user is represented as a TF-IDF vector, which is used as the input for an SVM classifier with the same setting as baseline 1.

5.6.3 Baseline 3: Word2Vec

Baseline 3 utilizes the word2vec embeddings to represent text. In this model, I used the entire collection of posts as the corpora and trained word vectors using the word2vec embeddings model (Mikolov, Chen, Corrado, & Dean, 2013). I used a CBOW model and set the dimension of the hidden layer to be 100to capture similarities among words. At last, each user is represented as a mean embedding vectors of the submissions that user authored/commented. I applied the SVM classifier on this representation to carry out the task.

5.6.4 Baseline 4: Label Propagation

Baseline 4 is the label propagation algorithm (Rossi, Lopes, & Rezende, 2014). The version of label propagation runs on a bipartite graph and it has a labeled set for only one type of the nodes in the graph. I directly applied it on the bipartite graph G_B to classify representative user. The initial labeled set is chosen based on degrees of user nodes, which is the same method used in the co-training model.

5.6.5 Baseline 5: Node2Vec

In Baseline 5, a Node2Vec (Grover & Leskovec, 2016b) model is applied to the data to learn the representations of the nodes. Then an SVM model is used to carry out the classification task. I set the dimensions of the embedding vectors to be 128. For the random walk process in the Node2Vec model, I tuned different values for the return parameter and the in-out parameter and report the best results here.

5.6.6 Baseline 6: Network&Attributes Embedding

In baseline 5, I applied another network embedding approach introduced in (Dave, Zhang, Chen, & Hasan,2018). This method learns hidden representation of vertices in a graph in combination with attributes of nodes, which allows us to incorporate textual information in the representation learning process. I used TF-IDF vectors of each user as user node attributes and set the dimensions of hidden layer to be 300, where150 for the graph and 150 to represent attributes. At last, I used the embedding vectors in the SVM classifier, which is the same as in previous baselines.

5.6.7 Performance Comparison

Overall, the co-training model performed better than all the baseline methods.

From the results of baselines 1 to 3, I see the performance of solely using language cues to classify users, which are not ideal despite trials of different methods. This is consistent with our observation that representative users' textual content is not that different from non-representative users in the long term in the dataset. Their online activities cover a wide range of topics beside their disability related content. They may use more disability related terms from time to time. However, the difference is not that

significant in the long run, especially from those non-representative online users who work in the healthcare domain.

Baselines 4-6 utilized graph-based algorithms to classifier online users. Baseline 4 showed very low recall. It is not surprising given the fact that the real-world social network is a very sparse and a disconnected graph. Hence, it is difficult to have a training-set that would propagate label information smoothly to all the unlabeled nodes. The values fluctuate depending on the choice of initial nodes. In Baseline 5, I used a representation learning method to learn embedding vectors first, which are later on used in a classification model. I found that this approach was not able to efficiently separate the target users in the graph. In baseline 6, I adopted a method that combines both network and textual information for each user. However, the performance showed almost the same result as baseline 5. From the results, I found that representation learning methods help mitigate the problem of sparsity in our classification task. However, it is not efficient in separating different dimensions of homophily, which leads to low performance in the experiment.

In conclusion, the results show that the co-training model generated the best results. By combining the label propagation algorithm with a NB classifier, the model mitigates the problem caused by the sparsity of the graph. It also shows that homophily is a viable way to identify online users with the same disabilities.

5.6.8 Efficiency of the Model

As aforementioned, the co-training model is semi-supervised, which would help mitigate the problem of manual annotation. In the experiment, I tried different sizes of training-set and showed the corresponding performances.

The results of different training-set sizes are summarized in Table 5.4. I set the training-set sizes to be 25%, 50%, and 75%. Since the training-set of the model is generated using random sampling, I run the model with each training-set size for 5 times and report the average performances here. The metrics are macro average across all classes. In Table 5.4, I observe that the co-training model archived a high precision (85%) and accuracy (81%) when the training-set is small (25% of the dataset). However, the recall (67%) is not ideal. With a larger training-set (75%), the recall of the model increased from 67% to 81%while the precision maintained the same level. The fact that the propose model can archive a high precision and accuracy with a small training-set can be very useful in identifying positive instances from a very large dataset.

Training-set size	**Precision**	**Recall**	**F1**	**Accuracy**
25%	0.85	0.67	0.72	0.81
50%	0.84	0.74	0.76	0.81
75%	0.84	0.81	0.80	0.82

Table 5.4: Performance comparison of the co-training model using different sizes of data as the training-set (macro avg.).

It is worth noting that the LPBG algorithm is essentially based on matrix multiplications. Hence, the speed of the model is largely determined by the choice of the auxiliary classifier. In our experiment, I applied a NB classifier. The only bottleneck is getting the representation of the text. I chose to use TFIDF, which is very efficient and was proven to be good enough in our experiment. Other auxiliary models and representations can be combined to solve classification tasks in other contexts.

The number of iterations m is another factor that would affect the training speed. Typically, the choice of m and K_{all} should be based on the size of the unlabeled set in the

dataset. In our model, since the LPBG can learn labels for all unlabeled instances in each run. The choice of m and K_{all} only affects the NB classifier. And the purpose is to let the NB classifier generate more positive instance, vertices with representative labels, after each iteration in order to provide the LPBG with more information. In our experiment, I set K_{all} to be 600 and m to be 10 because I observed that the NB classifier rarely generates positive instances after 10 iterations, which is due to the fact that the dataset is highly unbalanced in our experiment.

5.7 Phase Two Conclusion

The co-training model can help identify representative users with disabilities on public social media websites. Since this method leverages the homophily principle, it may generalize to identify members of other communities which exhibit homophily (for example, users with shared interests (Chang, Kumar, Gilbert, &Terveen, 2014)). Below, I discuss the implications and limitations of this approach.

The experiment showed good result when I applied this co-training model on a real-world dataset that has unbalance classes of training instances. The method is applicable to any social media website as long as its users' social network can be represented as a bipartite graph with a type of media as the communication channel (e.g., tweets, photo posts, etc.) Our approach specifically focuses on users who have self-disclosed their disability status online. I specifically examined this model on data from Reddit, an anonymous platform which may lend itself more readily to sensitive discussion and self-disclosure of potentially stigmatizing information. While our model likely does not identify many users with disabilities who are more private and do not disclose their disability status, this is an intentional limitation meant to preserve their

privacy and focus on users who are willing to openly discuss their disability and the ways it impacts their life. Of course, given the special design of the co-training model and our homophily hypothesis of the social network, the applicability of the model is limited to bipartite sparse graph with both graph and textual information. This limit is worth looking into in future work so the model can be applied in standardized datasets.

Chapter 6

Phase III: Data Visualization to Support Designers and Researchers

The work described in Chapters 4 and 5 can be used to identify a robust dataset of social media posts authored by representative users. However, in their original form, these posts are unstructured and may be difficult to make utilize as a resource to inform more representative design. In this chapter, I introduce a new data visualization tool that I designed, implemented, and evaluated to support scenario-based design for researchers in the accessibility domain using representative social media posts.

6.1 Motivation and Intuition of the Tool

A machine learning model to identify online users with disabilities can serve multiple purposes for accessibility researchers. First, by identifying online users with disabilities, researchers can collect online users' usernames and reach out to them for participant recruiting. This approach provides researchers with easy, low-cost assistance in recruiting representative users. Second, there is a large amount of user-generated content created by representative users already on social media websites. This data can be especially valuable to accessibility researchers and UX designers in the first phase of the design process (Preece, Rogers, & Sharp, 2002), where they research and build empathy with the population they are designing for.

Unfortunately, the amount of data available on online platforms is impractical to use for manual analysis and will keep growing, and this data is not structured or easily cataloged. As a result, the raw data is rarely used in the design process, and to my best knowledge, no existing work tried to explore the value of such data. This phase of my

dissertation work aims to create and test a data visualization tool to evaluate how this data could be meaningfully used by designers.

In Section 6.2, I describe the backend processes used to adapt our social media dataset for use in a data visualization tool. In order to facilitate the evaluation, I implemented a high-fidelity prototype of the data visualization tool based on real world data collected from social media websites. The frontend design and implementation of this tool are described in Sections 6.3 and 6.4. I design and carried out initial user studies to examine UX students' use of the prototype, which is analyzed and reported in Section 6.5.

6.2 Backend Design

As suggested in existing literature, there are many data visualization tools that focus on different types of data. The key for visualizing large datasets is to re-organize data to facilitate exploration while maintaining enough context (Steele & Iliinsky, 2010). I adopted the divide and conquer strategy (Roberts et al., 2014; Roschelle & Teasley, 1995), which means dividing the dataset into small parts for processing. Also, I adopt the PACT analysis pattern (Benyon & Macaulay, 2002) to facilitate exploration for researchers in the accessibility domain.

In order to allow data separation and PACT analysis, I first introduce the design of the backend of the visualization tool. The back end of the data visualization tool consists of five components, four of which are powered by machine learning or deep learning models. The overall pipeline is summarized in Figure 6.1.

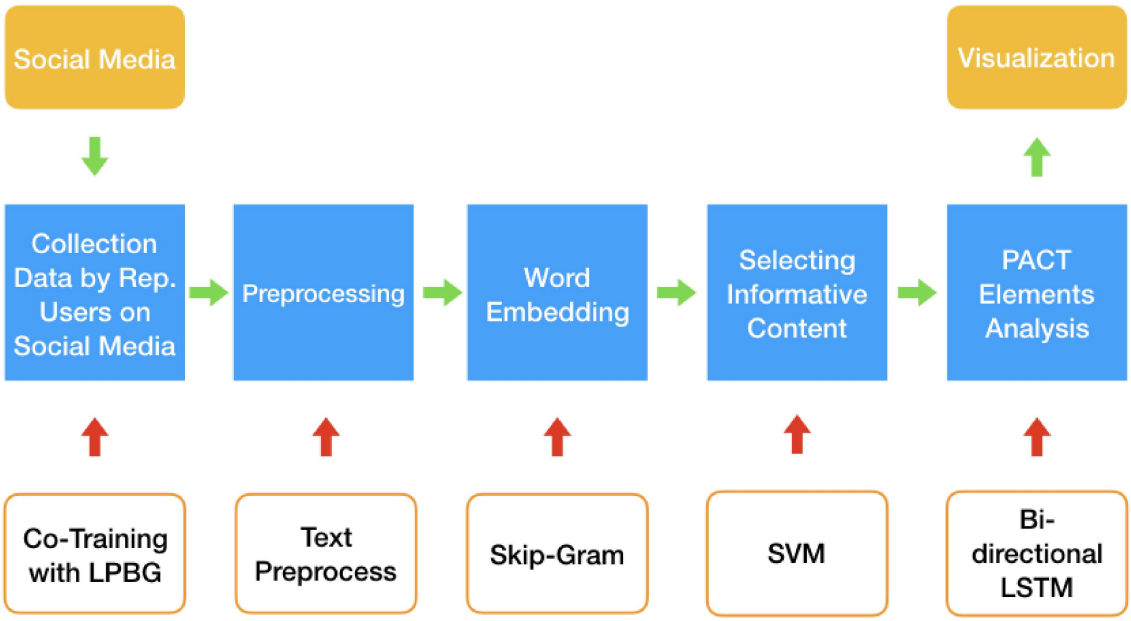

Figure 6.1: The pipeline of the data visualization tool

6.2.1 Social Media Data Input

The input of the visualization tool is raw data from social media websites. I apply the co-training model with LPBG, which was introduced in Chapter 5, to identify online representative users. After the classification task is finished, I collect the text content, which are the posts that are either authored or commented on by representative users. This dataset will be used as raw data for the next steps in the pipeline.

There are two things that are worth noting about the dataset generated by this set. First, the texts in the dataset are generated either by people who have a representative disability or people who are closely related to representative users, which makes it valuable in analyses. Second, the data are raw and mostly informal, natural language conversations about disability, without any potential bias that may happen while using traditional data collection methods such as interviews (Dell, Vaidyanathan, Medhi, Cutrell, & Thies, 2012).

6.2.2 Text Preprocessing

The text data derived from the last step will be passed to the text preprocessing step. In this step, an original copy of the text will be kept while another copy will be used to go through the following process. First, the text is cleaned for contractions and punctuation. Then it is tokenized into words with stop words removed. Based on the output tokens, the process generates a dictionary that assigns each token with a unique integer id. In addition, the frequency of appearance of each token is counted. The original text is transformed into a sequence of integers, which represents the tokens in the dictionary. This process is standard in natural language processing and produces required components for the next steps in the pipeline.

6.2.3 Word Embedding

In this section, the tokens from the previous steps are embedded into vectors using the Skip-Gram model (Mikolov et al., 2013). The model consists of one input layer, which takes one-hot vectors as input. Then there is an embedding layer with word embedding vectors, which is followed up with an output layer with a softmax activation function. For the purpose of performance, a noise contrastive estimation (NCE) loss is used instead of cross-entropy loss (Chen, Grangier, & Auli, 2016). I used 2,601,992 online posts from the study in Chapter 5 to train the Skip-Gram model.

In order to improve the learning outcome, I applied pre-trained GloVe embedding vectors[6] to initialize the Skip-Gram model. In this process, all converted text from the previous step will be used as the training set for the word embedding task. Overall, there were 930,762 unique tokens in the dictionary. The embedding size is set to 100. Also, the

[6] https://nlp.stanford.edu/projects/glove

sliding window length is set to 5 for negative sampling and the number sampled is set to 16 in the training process.

After this process, each token in the dictionary is represented as a one by hundred vector, and the information contained is passed on to the next two steps.

6.2.4 Selecting Informative Content

According to existing studies that focus on the quality of medical or health-related social media data (Denecke & Nejdl, 2009; Vlahovic, Wang, Kraut, & Levine, 2014), online data can be both informative (sharing medical information, terminology, or research) as well as experiential (describing an individual's own feelings and experiences living with a certain condition). Through examining online posts, studies (Denecke & Nejdl, 2009) defined online posts as either informative or affective, noting that informative posts are more likely to use formal medical terminology and affective posts are more likely to use adjectives. Examining the data generated by the representative users showed similar patterns - some posts are more informative, while others were used by online users to exchange emotional support. Hence, in this step, I designed a classifier to filter out informative texts for visualization.

For each of the posts, I tokenized the texts into words. For each of the token, the corresponding word embedding vector is looked up from the pretrained word vectors. An average pooling is carried out based on all the word vectors in a single post. Then the output is used in training a support vector machine (SVM) classifier for classification. I manually labeled 131 posts, in which 56 posts are affective and 75 posts are informative. I used this dataset to train the SVM model with a 70/30 split for training and testing. The

final classifier has a 77% precision and 87% recall for classifying informative online posts.

6.2.5 RNN Model for PACT Analysis

As described in Section 3.5.1, the PACT framework identifies people, activities, contexts, and technologies which may be relevant in performing design work and creating scenarios. I treat the PACT analysis as a Named Entity Recognition (NER) task. I adopt the model design structure from the deep learning model proposed in existing work (Collobert et al., 2011). The model is a recurrent neural network (RNN). It is a sequence to sequence model, which means that for each input (e.g., a word in the text), an output will be generated (e.g., a label indicating which PACT element it most likely is). In order to prevent vanishing gradients, I adopt the long short-term memory design (Hochreiter & Schmidhuber, 1997) in the network with an embedding layer to improve performance. An overview of the network is depicted in Figure 6.2.

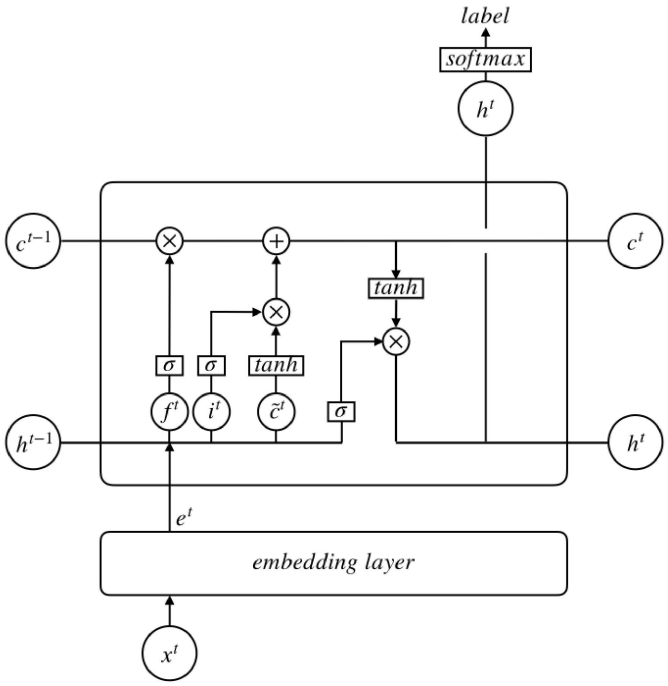

Figure 6.2: A LTSM unit with an embedding layer

The superscript t denotes step. At each time step, there are two inputs. First, there is the hidden status from the previous step, denoted as h^{t-1}. Second, the embedding vector, denoted as e^t, of the current word x^t in the text is passed into the model. There is one output at each step, which is the label of the input word.

Inside the LSTM unit, there are three values that are worth noting. The first is f^t, which controls forget in the unit. Then there is i^t, which controls the input of the new information. Finally, the new input is denoted as $\tilde{c}^t$.

$$f^t = \sigma(w_1 < h^{t-1}, e^{t-1} + b_1)$$

$$i^t = \sigma(w_2 < h^{t-1}, e^{t-1} + b_2)$$

$$\tilde{c}^t = tanh(w_3 < h^{t-1}, e^{t-1} > + b_3$$

$$c^t = f^t \times c^{t-1} + i^t \times \tilde{c}^t$$

These three values are calculated as shown in the equations above. σ denotes the sigmoid function. w_i and b_i denote different parameters and bias terms in calculations. $tanh$ denotes the tanh function. At each step, the information in the memory cell, denoted as c^{t-1}, is updated by f^t, it, and $\tilde{c}^t$ as the last equation. The updated cell value c^t is passed onto the next step. Also, the hidden status h^{t-1} is updated as following:

$$h^t = tanh(c^t) \times \sigma(w_4 < h^{t-1}, e^{t-1} + b_4 >)$$

After h^t is computed, the value is used in for two purposes. First, it is passed onto the next step as the hidden status from the previous step. Second, it is passed through a softmax function to derive a label for the current input word x^t.

As mentioned in Section 3.5.1, a sequence-to-sequence model uses previous information to learn and predict the current input. The underlying assumption is that existing and past observations are sufficient to predict the current outcome. However, in

natural language processing, it is not always the case. For example, if I input a sentence such as “Dan is a dog” in a NER classifier, the words after “Dan” contains important information to correctly label “Dan”. Thus, based on the LSTM model, I set the RNN model to be bidirectional to improve classification performance.

To train the bidirectional long short-term memory RNN, I labeled 400 sentences using the format shown in Table 6.1. First, the text is tokenized into sentences. Each sentence is an input sequence. Then, each sentence is tokenized into words and punctuation. Each of the tokens has a corresponding label. There are five different types of labels, which are “person”, “context”, “technology”, “activity”, and “O”. The “O” tag stands for no label.

Token	**Label**
Hey	O
,	O
I	person
‘ve	O
gone	O
through	O
a	O
knee	context
injury	context
(	O
missing	context
a	context

knee	context
ligament	context
now	O
,	O
will	O
probably	O
need	O
replacement	technology
down	O
the	O
road	O
…	O
.not	O
why	O
I	O
‘m	O
on	O
this	O
sub	O
,	O
just	O
to	O
be	O

clear	O
)	O
and	O
had	O
some	O
thoughts/questions	O
for	O
you	O
.	O

Table 6.1: Example of manual annotation for the bidirectional LSTM RNN training data.

In the training process, the label sequences are used as input. Each token is first transformed as an embedding vector from the results of the previous model. Unknown tokens are represented using the average word embedding vectors. This embedding layer is not trainable.

After tuning the parameters and hyperparameters of the model, the final model has the following setup. The LSTM cell has 200 neurons. The embedding size is 100. To prevent over-fitting, I adopted the dropout strategy. The dropout rate is 0.5 in each step, and the recurrent dropout rate of the RNN is set to be 0.25. 90% of the labeled data is used as the training data and the remaining 10% is used as the development set to evaluate the performance.

The model is trained with 1574 labeled sequences from online posts. The training accuracy of the model with 1000 epoch is 99.64%. The average accuracy of the development set is 87.60%.

6.3 Frontend Design

I created a frontend application to visualize the data generated in Section 6.2 for researchers and designers. I introduce the design and features of the frontend application in this section.

6.3.1 User Interface Design

The user interface of the data visualization tool is depicted in Figure 6.3. The interface has three parts. There is a control panel in the top area. This panel contains three control features and a help button. Beneath the control panel, there are two panels side by side. The right-side panel contains free text, and the left side panel contains an interactive graph, which I call a force graph.

Figure 6.3: User interface of the data visualization tool.

6.3.2 The Free Text Panel

The right-side panel contains the free text from posts that are either authored or commented by representative online users on Reddit.com. Each box in the panel contains a sentence from the posts. There are colored boxes for some of the words in the text. These were generated by the bidirectional LSTM RNN model from the backend. In other words, the colored boxes represent what the PACT elements this word belongs to. Red denotes context related words. Green denotes person related words. Blue denotes technology related words. Orange denotes activity related words. I named these color boxes as PACT highlighting. The highlighting is powered by the backend NER model, which can be updated when more annotations are available.

6.3.3 The Interactive Force Graph

The interactive force graph in the left panel serves as a summary of the text data in the free text panel. The graph consists of multiple nodes with different sizes and colors. Each node represents a PACT element from the text. The color coding follows the same pattern in the free text panel. For example, a red node represents a context related word. The size of the node represents the frequency of appearance of this word in the text. The larger the node is, the more frequent it appears in the text panel. The nodes in the graph are not static. They are connected to each other. I call this a force graph because whenever you use the mouse to drag a node, other nodes that are connected to the node will be dragged as well. The strength of the force between two nodes is based on the euclidean distance between the embedding vectors of the two nodes. The larger the distance is between two nodes, the weaker their connecting force will be. For each of the nodes, they also have a repelling force from other nodes to present node overlapping.

6.3.4 Show Words

A toggle button is located in the middle of the control panel for users to show or hide the corresponding labels of the nodes. When the button is turned, the visualization changes as shown in Figure 6.4.

Besides dragging a node, double click is also available. A user can double click on a node to trigger a quick search, which will lead to changes in the text panel on the right side. The panel will only show the text that contains the word of that node. In Figure 6.4, the word labels of nodes are clustered and overlapping. Users can click and drag a node to move the node and the word label together to make it isolated and clear to read. After releasing, the node will bounce back to its original relative position.

Figure 6.4: User interface with show words turned on.

6.3.5 Cut Words

As aforementioned, the words have link force to tie them together. However, while the link force shows the relationship among different words, it can sometimes make the graph too clustered to find interesting patterns.

I designed a cut words feature in this visualization tool to help improve the interaction of the graph. This feature is controlled by a sliding bar in the control panel. A user can click and drag to adjust the slider to change the value from 1% to 99%, which represent the normalized Euclidean distance. When the slider is set to the left end of the bar, any pairs of word with a distance that is smaller than 1 will be affected by the link force. The nodes in the graph are clustered as shown in Figure 6.5. On the other hand, if a user sets the slider to the other end, then only pairs of nodes with a euclidean distance close to 0 will have a link force. Hence, all nodes will be separated as shown in Figure 6.6.

By adjusting the slider to a value in the middle range, a user could cut off the words that are not strongly linked together and find clusters that have strong connections as depicted in Figure 6.7. The cutoff will keep the link force of pairs of nodes that have euclidean distances that are lower than 0.9. By doing this, a user can identify PACT elements that are strongly connected together. One example is shown in Figure 6.8. After setting the cut off value to 0.9, the three words “toe”, “bone”, and “doctor” are still together. This cluster shows that these three words appear frequently together in the text. A user could adjust the value of cutoff words to find useful clusters.

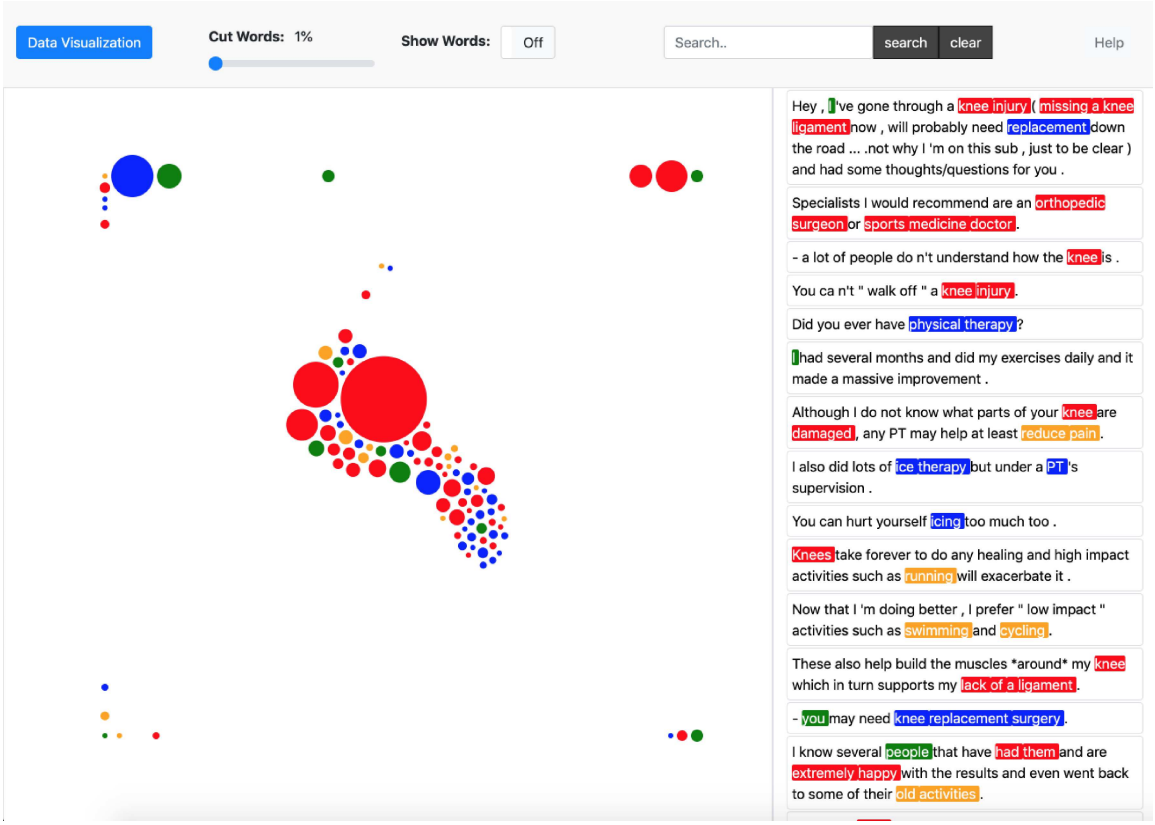

Figure 6.5: User interface with cut words set to 1%.

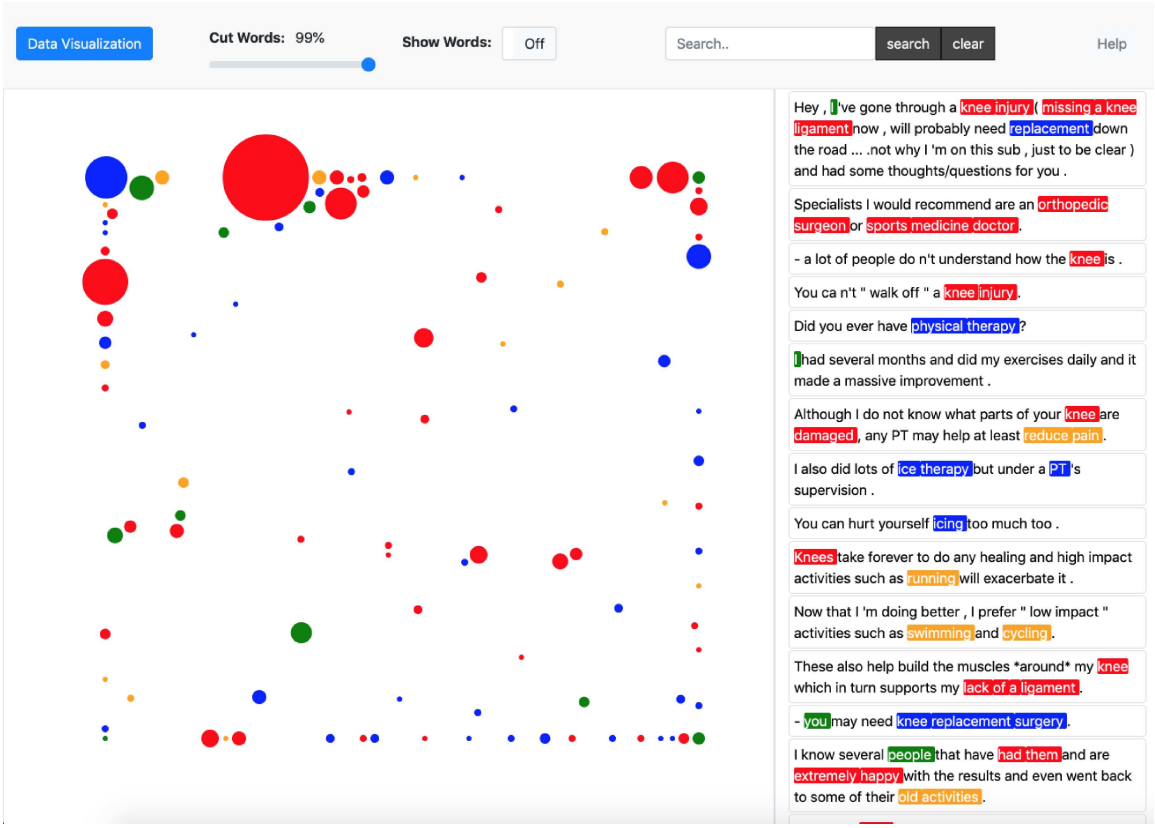

Figure 6.6: User interface with cut words set to 99%.

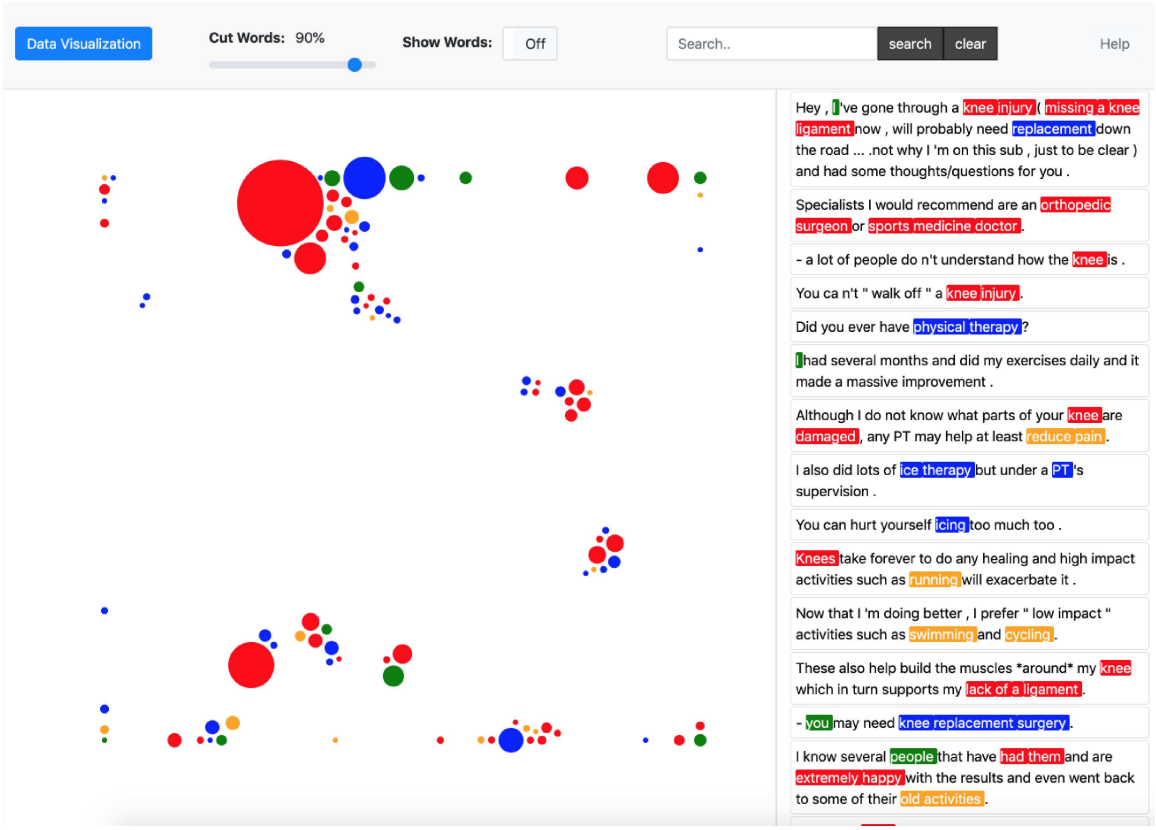

Figure 6.7: User interface with cut words set to 90%.

Figure 6.8: Example of a cluster based on cut words set to 90%.

6.3.6 Search

The prototype provides a basic search function. A user can search a single word at each time. The results will show texts that contain the search word. An example of the search function is depicted in Figure 6.9.

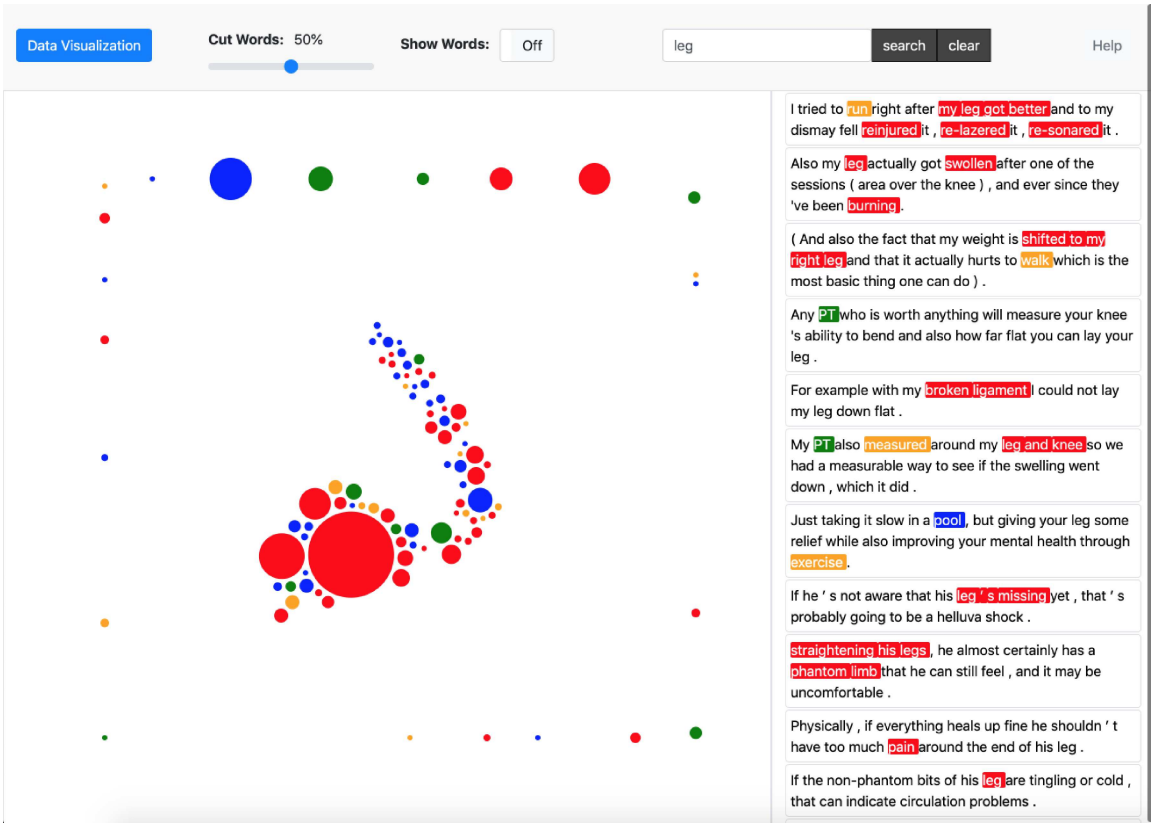

Figure 6.9: Example of search results with query “leg”.

6.4 Implementation of the Data Visualization Tool

The data visualization is implemented using a browser/server structure. The frontend application runs in web browsers. The backend runs on a web server. The backend models are implemented using python with several machine learning and deep learning libraries that include NLTK[7], Scikit-learn[8], and Tensorflow[9]. The output data is passed onto the web browser to generate the visualization. A model-view-controller paradigm is used to create the backend application. The data model and the controller handle the data from the machine learning/deep learning models and generate a view, which is passed on to the browser for visualization in the JSON format.

[7]http://www.nltk.org
[8] http://scikit-learn.github.io/stable
[9] https://www.tensorflow.org

The front end is implemented using HTML, CSS, Javascript, D3[10], bootstrap[11], and jQuery[12].

6.5 Pilot Study of the Data Visualization Tool

In order to evaluate the data visualization tool, I carried out a qualitative user study. In this section, I introduce the method, data, and results of the user study.

6.5.1 Method

I designed the evaluation to be an exploratory study of how HCI practitioners use the tool through two design sessions, which are followed up with a semi-structured interview. I ask each participant to do a task, which is to find a problem related to amputees. Given this prompt, each participant tried to operate a computer with Internet connection. The participants were asked to think aloud during this process. I repeated this task for two sessions. In the first session, the participants will use an online search engine to explore and find information. In the second session, the participants will have access to the data visualization tool to perform the task. After each session, I conduct a semi-supervised interview to collection participants' feedback on the process. This allows me to understand their think process and search strategies during problem finding. Each task session lasts for 10 minutes while each interview session lasts for 5 minutes. After the two sessions and two interviews are completed, I conduct more interview to collect participants reflection on how their process changed in the two sessions.

[10] https://d3js.org
[11] http://www.bootstraptoggle.com
[12] https://jquery.com

6.5.2 Participants and Data Collection

I conducted an exploratory study with 4 participants. The participants are all Ph.D. students in human- computer interaction major with at least three years of experience in the HCI domain, recruited by convenient sampling. For each participant, I collected the following data: 1) observation logs of the problem identification task, 2) digital recordings of the interview, 3) written notes generated by participants during their task session.

The semi-supervised interviews followed the following procedure:

INTRODUCTION:

Thank you for your time. We want to evaluate a visualization tool to see if it can help problem finding for designers and researchers. This process will take 30 minutes in total. We would like to ask you to form a design question related to amputees using different approaches.

SEARCH ENGINE ACTIVITY:

First, we would like you to find a design problem and form a design question using a search engine. The topic is about amputees. You can use any techniques along the way to specify the design problem (creating personas, scenarios, etc). Please take no more than 10 minutes.

REFLECTION ON SEARCH ENGINE ACTIVITY:

Would you reflect a little bit on the process of using the search engine to find this information, and elaborate on your thoughts? What difficulties did you encounter? What kind of help do you want?

DATA VIZ ACTIVITY:

Second, we would like you to do the same task using a newly designed data visualization tool in addition to the search engine (give a tour of the tool). You will be working in the same domain with the same time limit, and you can generate as many results as you like.

DATA VIZ ACTIVITY REFLECTION:

Would you reflect a little bit on the process of using the data visualization tool and elaborate on your thoughts? What difficulties did you encounter? What kind of help do you want?

OVERALL COMPARISON:

Would you reflect the difference in using the search engine and the viz tool? Would you compare the approaches and give some pros and cons for both? When in the design process would you want to use the search engine to find information, and when would you want to use the data viz tool?

COMMENTS:

Do you have any thoughts, comments, and suggestions for the design of the visualization tool? What do you want to see that is different in the future?

The semi-structured interview focused on (i) the thinking process and strategies each participant adopted; (ii) how the process and strategies changed after participants were provided with the data visualization tool, and (iii) opinions the participants formed about the tool.

I transcribed parts of the interview recordings and analyzed the transcriptions and the observation logs to focus on several aspects. First, I observe how participants interact

with the data visualization tool. Second, I pay special attention to the behavioral changes that appeared while using the tool. Third, I identified difficulties they encountered during the process.

6.5.3 Results

I analyzed the data using an affinity diagramming process (Beyer & Holtzblatt, 1999) to find common themes in the participants' use of the tool. The results are introduced in this section.

General Patterns in Problem Finding Process

All four participants adopted a very similar strategy in the initial task session (before using the data visualization tool). All participants started their search with an intuitive query using a general search engine. Some examples of the queries are "problem faced by amputees in public space" and "amputee organizations". The queries are mostly based on life or research experience. The purpose is mainly to find useful information regarding the given design task. All four participants quickly modified the initial query and tried to find more specific data. However, they all failed to find useful contextual information they need to get a quick overview of the field with a general search engine, since it is difficult to generate informative queries that could narrow down the search results. All participants started searching in specific search services such as Google Scholar, ACM Digital Library, Wikipedia, and organization websites related to amputees. In this process, participants tried to find useful information by quickly reading through the snippets of research results until the session time is up.

Difficulties in Problem Finding

Two difficulties were consistently mentioned by all 4 participants. The first is the difficulty of forming helpful queries. Participants all mentioned that the queries they used did not provide good results. They have an intuition of what to look for but don't know the right language. The second difficulty is the amount of helpful information and the mental load to process them in the search results. All participants reported that, due to the difficulty of forming useful queries, the results in the search engine do not contain a lot of useful information. They all wanted to look for a kind of overview of the problem domain (e.g., a literature review paper, a descriptive statistic of from authorities).

Interactions with the Data Visualization Tool

During the second session, in which the data visualization tool is introduced to participants. I observed how participants interacted with the tool. All four participants had an easy learning process in using the tool. They adopted the features fast. In the second session, participants spend more time interacting with the data visualization tool. The interactive pattern has two styles. Two participants used an exploratory style. They used the force graph to find most frequent PACT elements. Following which, they started to look at relevant nodes. For most of the nodes, they searched for the corresponding text and read through the context. The other two participants adopted a search first style. They used the search function in the tool to search for a query they are interested in and used the force graph and the text to look for related content. In both of the styles, an iterative search pattern has been observed among all participants. They started forming new queries faster than the previous session. Some of them used the general search engine with the new query to help find more context information.

Comparisons Between Two Sessions

By examining the results, the following differences are observed between the two sessions from the participants. First, participants showed different time allocation in the tasks. While using a traditional search engine, all participants used their time trying to narrow down search queries through different approaches (e.g., use different search tools, check different websites). With the visualization tool, all participants spent their time in exploring related contents instead of forming useful queries. Second, while using the data visualization tool, participants started to look at data from social media, which was not paid attention to while using a traditional search engine. Participants reported that the tool helped information foraging by mitigating the necessary mental load to process the data. Third, three out of four participants were able to narrow down their search scope for problem finding into a specific topic with the data visualization tool. They were able to scale down their search into 3 to 4 different queries and look for relevant data.

Overall Feedback

The data visualization tool received overall positive feedback from participants for several aspects: First, the tool helps search by providing quick context information. Users can explore potential search queries in the force graph while have access to the contextual information, online posts in this case, from which the queries were extracted. Second, the design of the tool fits some participants mental model more than general search engines. Third, it is very helpful in understanding end users' experience and problems.

Based on individual experience, participants also provided suggestions for improvements of the tool. One common requirement is to better support iterative search. Participants want to search not only the text but also the force graph. They would like to conduct a

search using multiple queries in an iterative manner, which means they want to be able to refine the search results by being allowed to select a part of the force graph and zoom in to conduct deeper search. In addition to iterative search, another feature that all participants mentioned is to be able to get more context data in the text panel. They would like to have the ability to see the source of the conversation. Third, they would like to see most elements other than PACT elements such as demographic information.

6.6 Phase Three Conclusion

In this study, I found that the process of problem finding for researcher/designers includes actions of finding useful information. This process is, on some level, very similar to exploratory search (Marchionini, 2006) in the information retrieval domain. Researchers try to satisfy their information need by applying different search strategies, and they stop until their information gain reaches a satisfactory level. However, such a task is difficult when exploring a new topic in the accessibility domain. First, search engines are very good at getting target information when the searcher has a clear search goal in mind. But forming a good search query requires domain knowledge, which is a barrier for newcomers to the domain. Second, as powerful as modern search engines are, researchers/designers still need to verify the usefulness and credibility of the search results. Reliable data is important to provide that information. By implementing and evaluating the data visualization tool, I found that the tool has the potential to support quick domain knowledge gain and query exploration. On a higher level, the tool brings what a search engine is lacking, which is browsing. Browsing can very powerful for information seeking in an exploratory search. Instead of narrowing down the information,

which a search engine does, browsing expands the information gain for serendipitous findings.

Chapter 7

Conclusion

7.1 Discussion and Conclusion

In my thesis, I proposed a new direction to empower accessibility researchers by using data from social media websites. The task is challenging and never studies before. I first designed a new model to classifying online users with disabilities for data collection and participants recruiting. Then, I design, implemented, and evaluated a data visualization tool to facilitate problem finding for researchers/designers in this domain. Through this process, I devised a new label propagation algorithm that works on a bipartite graph. I proposed a new co-training model to help improve label propagation on sparse graphs. I also trained word embedding for a collection of words related to disabilities. And lastly, I design a data visualization tool based on the PACT analysis for researchers.

This tool can be used independently for accessibility researchers as an augment to existing search engines to help problem finding in the formative stage of a research project. It also can be embedded in existing qualitative analysis tools to help researchers find useful information in social media data. For online users, it is also possible to embed the tool as a browser plugin for them to find useful information that pertains to their conditions.

7.2 Implications and Future Work

All these contributions I made pile up to a potential future solution for improving accessibility studies. As my work is only a beginning step towards exploring the full

potential of social media data's application in the accessibility domain, it pointed out directions for future work.

First, the potential of social media data is not only limited to problem finding or participants recruiting. At this stage, there lacks a comprehensive approach, given the complexity of social media data, to accomplish other tasks. However, a more advanced model could be designed to provide other information that is useful for researchers such as demographic composition and automatic problem identification.

Second, there is more research to conduct as to how information could be used to augment researchers' work. The sheer amount of social media data is out of human's capacity for processing. Data visualizations powered by machine learning models provide a way to augment the process. But more work is needed to address problems such as how to balance being concise and provide enough context. How to make tools unbiased and acceptable to researchers.

Last but not least, more work is required to address the problem of privacy. Since social media data can be private and sensitive, abuse of the data could lead to potential ramifications. The need to incorporate more work from a cybersecurity perspective is just as important.

Appendix

Appendix A

1. Python Code for the LPBG Algorithm

Repository URL: https://github.com/xing-yu/label_propagation_on_bipartite_graph

2. Code for the Data Visualization Tool

Repository URL: https://github.com/xing-yu/viz

Demo of the Visualization Tool

Demo URL: https://www.youtube.com/watch?v=z-4clWa9Dkfeature = youtu.be

Reference

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, *6*(2), 9.

Antheunis, M. L., Tates, K., & Nieboer, T. E. (2013). Patients' and health professionals' use of social media in health care: motives, barriers and expectations. *Patient education and counseling*, *92*(3), 426–431.

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily- driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, *106*(51), 21544–21549.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, *12*(9).

Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., . . . Aly, M. (2008). Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th international conference on world wide web* (pp. 895–904).

Benyon, D., & Macaulay, C. (2002). Scenarios and the hci-se design problem. *Interacting with computers*, *14*(4), 397–405.

Beyer, H., & Holtzblatt, K. (1999). Contextual design. *interactions*, *6*(1), 32–42.

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1997). Nymble: a high-performance learning name-finder. In *Proceedings of the fifth conference on applied natural language processing* (pp. 194–201).

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

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Brady, E. L., Zhong, Y., Morris, M. R., & Bigham, J. P. (2013). Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 conference on computer supported cooperative work* (pp. 1225–1236).

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1301–1309).

Calle, M. L., & Urrea, V. (2011). Letter to the editor: stability of random forest importance measures. *Briefings in bioinformatics*, *12*(1), 86–89.

Chang, S., Kumar, V., Gilbert, E., & Terveen, L. G. (2014). Specialization, homophily, and gender in a social curation site: findings from pinterest. In *Proceedings of the 17th acm conference on computer supported cooperative work & social computing* (pp. 674–686).

Chen, W., Grangier, D., & Auli, M. (2016). Strategies for training large vocabulary neural language models. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1975–1985).

Chittaro, L., & Dal Cin, P. (2002). Evaluating interface design choices on wap phones: Navigation and selection. *Personal and Ubiquitous Computing*, *6*(4), 237–244.

Chittaro, L., & De Marco, L. (2004). Driver distraction caused by mobile devices: studying and reducing safety risks. In *Proceedings of the 1st int'l workshop mobile technologies and health: Benefits and risks (udine, italy, 2004)*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, *12*(Aug), 2493–2537.

Compeau, P. E., Pevzner, P. A., & Tesler, G. (2011). How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, *29*(11), 987–991.

Corley, C. D., Cook, D. J., Mikler, A. R., & Singh, K. P. (2010). Text and structural data mining of influenza mentions in web and social media. *International journal of environmental research and public health*, *7*(2), 596–615.

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM*, *39*(1), 80–91.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., & Suri, S. (2008). Feedback effects between simi- larity and social influence in online communities. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 160–168).

Dave, V. S., Zhang, B., Chen, P.-Y., & Hasan, M. A. (2018). Neural-brane: Neural bayesian personalized ranking for attributed network embedding. *arXiv preprint arXiv:1804.08774*.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *ICWSM*, *13*, 1–10.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 2098–2110).

Dee, M., & Hanson, V. L. (2014). A large user pool for accessibility research with representative users. In *Proceedings of the 16th international acm sigaccess conference on computers & accessibility* (pp. 35–42).

Dell, N., Vaidyanathan, V., Medhi, I., Cutrell, E., & Thies, W. (2012). Yours is better!: participant response bias in hci. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1321–1330).

Denecke, K., & Nejdl, W. (2009). How valuable is medical social media data? content analysis of the medical web. *Information Sciences*, *179*(12), 1870–1880.

Do ̈rk, M., Gruen, D., Williamson, C., & Carpendale, S. (2010). A visual backchannel for large-scale events. *Visualization and Computer Graphics, IEEE Transactions on*, *16*(6), 1129–1138.

Dou, W., Wang, X., Skau, D., Ribarsky, W., & Zhou, M. X. (2012). Leadline: Interactive visual analysis of text data through event identification and exploration. In *Visual analytics science and technology (vast), 2012 ieee conference on* (pp. 93–102).

Fischer, C. S. (1982). *To dwell among friends: Personal networks in town and city*. University of chicago Press.

Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-scale sentiment analysis for news and blogs. *ICWSM*, *7*(21), 219–222.

Grover, A., & Leskovec, J. (2016a). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 855–864).

Grover, A., & Leskovec, J. (2016b). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.*

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464–472.

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

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems* (pp. 856–864).

Kleinberg, J. M. (2007). Challenges in mining social network data: processes, privacy, and paradoxes. In *Proceedings of the 13th acm sigkdd international conference on knowledge discovery and data mining* (pp. 4–5).

Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, *11*, 538–541.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web* (pp. 591–600).

Langville, A. N., & Meyer, C. D. (2011). *Google's pagerank and beyond: The science of search engine rankings*. Princeton University Press.

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

Light, R. P., Polley, D. E., & Bo¨rner, K. (2014). Open data and open code for big science of science studies. *Scientometrics*, *101*(2), 1535–1551.

Ma, X., Hancock, J., & Naaman, M. (2016). Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 3857–3869).

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, *49*(4), 41–46.

Maynard, D., Bontcheva, K., & Rout, D. (2012). Challenges in developing opinion mining tools for social media. *Proceedings of the@ NLP can u tag# usergeneratedcontent*, 15–22.

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

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, *90*(5), 862.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Seventh international conference on learning representations (iclr).*

Morris, M. R., Perkins, A., Yao, C., Bahram, S., Bigham, J. P., & Kane, S. K. (2016). "with most of it being pictures now, i rarely use it": Understanding twitter's evolving accessibility to blind users. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 5506–5516).

Naaman, M. (2012). Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications. *Multimedia Tools and Applications*, *56*(1), 9–34.

Pasupathi, M. (2007). Telling and the remembered self: Linguistic differences in memories for previously disclosed and previously undisclosed events. *Memory*, *15*(3), 258–270.

Paul, M. J., & Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. *ICWSM*, *20*, 265–272.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, *77*(6), 1296.

Petrie, H., Hamilton, F., King, N., & Pavan, P. (2006). Remote usability evaluations with disabled people. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1133–1141).

Piskorski, J., & Yangarber, R. (2013). Information extraction: past, present and future. In *Multi-source, multilingual information extraction and summarization* (pp. 23–49). Springer.

Plaisant, C. (2004). The challenge of information visualization evaluation. In *Proceedings of the working conference on advanced visual interfaces* (pp. 109–116).

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, *10*(2), 191–218.

Preece, J., Rogers, Y., & Sharp, H. (2002). *Interaction design: Beyond human-computer interaction*. Wiley. Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on search and mining user-generated contents* (pp. 37–44).

Riloff, E., et al. (1993). Automatically constructing a dictionary for information extraction tasks. In *Aaai* (pp. 811–816).

Roberts, J., Lyons, L., Cafaro, F., & Eydt, R. (2014). Interpreting data from within: supporting humandata interaction in museum exhibits through perspective taking. In *Proceedings of the 2014 conference on interaction design and children* (pp. 7–16).

Robertson, G., Ebert, D., Eick, S., Keim, D., & Joy, K. (2009). Scale and complexity in visual analytics. *Information Visualization*, *8*(4), 247–253.

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning* (pp. 69–97).

Rossi, R. G., Lopes, A. A., & Rezende, S. O. (2014). A parameter-free label propagation algorithm using bipartite heterogeneous networks for text classification. In

Proceedings of the 29th annual acm symposium on applied computing (pp. 79–84).

Rosson, M. B., & Carroll, J. M. (2009). Scenario based design. *Human-computer interaction. Boca Raton, FL*, 145–162.

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121–1133.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web* (pp. 851–860).

Saket, B., Scheidegger, C., Kobourov, S. G., & Bo¨rner, K. (2015). Map-based visualizations increase recall accuracy of data. In *Computer graphics forum* (Vol. 34, pp. 441–450).

Sandri, M., & Zuccolotto, P. (2012). A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*.

Satuluri, V. (n.d.). Scalable graph clustering using stochastic flows.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *Aaai spring symposium: Computational approaches to analyzing weblogs* (Vol. 6, pp. 199–205).

Schmidt, C. W. (2012). Using social media to predict and track disease outbreaks. *Environmental health perspectives*, *120*(1), A31.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., . . . others (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791.

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

Sears, A., Karat, C.-M., Oseitutu, K., Karimullah, A., & Feng, J. (2001). Productivity, satisfaction, and interaction strategies of individuals with spinal cord injuries and traditional users interacting with speech recognition software. *Universal Access in the information Society*, *1*(1), 4–15.

Shen, Z., & Ma, K.-L. (2008). Mobivis: A visualization system for exploring mobile data. In *Visualization symposium, 2008. pacificvis'08. ieee pacific* (pp. 175–182).

Shoro, A. G., & Soomro, T. R. (2015). Big data analysis: Apache spark perspective. *Global Journal of Computer Science and Technology*, *15*(1).

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The hadoop distributed file system. In *Mass storage systems and technologies (msst), 2010 ieee 26th symposium on* (pp. 1–10).

Snoek, C. G., Worring, M., & Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual acm international conference on multimedia* (pp. 399–402).

Steele, J., & Iliinsky, N. (2010). *Beautiful visualization: Looking at data through the eyes of experts*. " O'Reilly Media, Inc.".

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), 1.

Szummer, M., & Jaakkola, T. (2001). Partially labeled classification with markov random walks. In *nips* (Vol. 14).

Talukdar, P. P. (2009). Topics in graph construction for semi-supervised learning.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tong, H., He, J., Li, M., Zhang, C., & Ma, W.-Y. (2005). Graph based multi-modality learning. In *Proceedings of the 13th annual acm international conference on multimedia* (pp. 862–871).

Tsandilas, T., Bezerianos, A., & Jacob, T. (2015). Sketchsliders: Sketching widgets for visual exploration on wall displays. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 3255–3264).

Vlahovic, T. A., Wang, Y.-C., Kraut, R. E., & Levine, J. M. (2014). Support matching and satisfaction in an online breast cancer support community. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1625–1634).

Voykinska, V., Azenkot, S., Wu, S., & Leshed, G. (2016). How blind people interact with visual content on social networking services. In *Proceedings of the 19th acm*

conference on computer-supported cooperative work & social computing (pp. 1584–1595).

Walker, B. N., & Mauney, L. M. (2010). Universal design of auditory graphs: A comparison of sonification mappings for visually impaired and sighted listeners. *ACM Transactions on Accessible Computing (TACCESS)*, *2*(3), 12.

Walker, M., Thornton, L., De Choudhury, M., Teevan, J., Bulik, C. M., Levinson, C. A., & Zerwas, S. (2015). Facebook use and disordered eating in college-aged women. *Journal of Adolescent Health*, *57*(2), 157–163.

Wang, F., Zhang, C., Shen, H. C., & Wang, J. (2006). Semi-supervised classification using linear neighborhood propagation. In *Computer vision and pattern recognition, 2006 ieee computer society conference on* (Vol. 1, pp. 160–167).

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third acm international conference on web search and data mining* (pp. 261–270).

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347–354).

Wu, S., & Adamic, L. A. (2014). Visually impaired users on an online social network. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 3133–3142).

Yu, X., & Brady, E. (2017). Understanding and classifying online amputee users on reddit. In *Advances in social networks analysis and mining (asonam), 2017*

international conference on Advances in Social Networks Analysis and Mining 2017 (pp. 17-22). ACM.

Zhu, X., & Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation. Zoidi, O., Fotiadou, E., Nikolaidis, N., & Pitas, I. (2015). Graph-based label propagation in digital media: A review. *ACM Computing Surveys (CSUR)*, *47*(3), 48.

Curriculum Vitae

Xing Yu

Education

- PHD. Major: Informatics, Minor: Computer Science. Indiana University, IUPUI. 2020.
- B.S. Major: Psychology. Zhejiang University, Hangzhou, China, 2008.

Research Interests

Data mining; Social Network; Machine Learning; HCI

Awards

- Google Student Award | W4A Conference 2016
- Funded Participant | HCIC Conference 2017

Experience

- Graduate Research Assistant at Indiana University | Aug 2014 – July 2019
- Director at Haina Corporation | 2009-2013
- Intern at Hewlett-Packard | 2008-2009
- Intern at Wasu | 2008

Teaching

- Instructor, Introduction to Informatics (I101) | Fall 2017, Fall 2018
- TA, Introduction to HCI (I270) | Spring 2017
- TA, Social Computing (H565) | Fall 2016

Presentations

- Doctoral Consortium: "Using Data from Social Media Websites to Inspire the Design of Assistive Technology" | Apr 2016 | Proceeding of the 13th Web for All Conference | Montreal Quebec, Canada.
- Poster: "Using Social Media Websites to Inform the Design of Assistive Technology" | June 2017 | Human Computer Interaction Consortium (HCIC) Conference 2017 | Pajaro Dunes Resort, Watsonville, CA, USA.
- Paper Presentation: "Understanding and Classifying Online Users with Physical Disabilities on Reddit" Aug 2017 | The 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining | Sydney, Australia

Publications

Yu, X., Chakraborty, S., & Brady, E. (2019, July). A Co-Training Model with Label Propagation on a Bipartite Graph to Identify Online Users with Disabilities. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 13, No. 01, pp. 667-670).

Yu, X. Brady, E. Understanding and Classifying Online Users with Physical Disabilities on Reddit. Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (New York, NY, USA), ASONAM '17, ACM, 2017, pp. 17–22.

Niu, X. Li, C. and Yu, X.(2017), Predictive analytics of e-commerce search behavior for conversion. In Proceedings of the 22nd Americas Conference on Information Systems.

Xia, T., Yu, X., Gao, Z., Gu Y., Liu, X. Internal/External Information Access and Information Diffusion in Social Media. In Proceedings of the 2017 iConference 2(2017), 129-133.

Jia, Y.,Liu, Y., Yu, X., Voida. S., Designing Leaderboards for Gamification: Perceived Differences Based on User Ranking, Application Domain, and Personality Traits, CHI2017, ACM.

Niu, X., Yu, X. (2016). Exploring Customers' Search Behavior on a Large E-Commerce Website. In Proceedings of the 22nd Americas Conference on Information Systems.

Liu, X., Yu, X., Gao, Z., Xia, T., Bollen, J. (2016, July). Comparing Community-based Information Adoption and Diffusion Across Different Microblogging Sites. In Proceedings of the 27th ACM Conference on Hypertext and Social Media (pp. 103-112). ACM.

Yu, X. (2016, April). Using data from social media websites to inspire the design of assistive technology. In Proceedings of the 13th Web for All Conference (p. 38). ACM.