GEOGRAPHICAL ENGAGEMENT AND CHURN PREDICTION IN LOCATION-BASED SOCIAL NETWORKS

by

YOUNG DAE KWON

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Master of Philosophy
in Computer Science and Engineering

January 2020, Hong Kong

Copyright © by Young Dae Kwon 2020
Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

YOUNG DAE KWON

January 2020
GEOGRAPHICAL ENGAGEMENT AND CHURN PREDICTION IN LOCATION-BASED SOCIAL NETWORKS

by

YOUNG DAE KWON

This is to certify that I have examined the above M.Phil. thesis and have found that it is complete and satisfactory in all respects, and that any and all revisions required by the thesis examination committee have been made.

Prof. Pan Hui, Thesis Supervisor

Prof. Dit-Yan Yeung, Head of Department

Department of Computer Science and Engineering

January 17, 2020
ACKNOWLEDGMENTS

I would never have completed this work without help from many people. First of all, I thank my advisors, Professor Pan Hui and Professor Raymond Chi-Wing Wong, for their years of mentoring, advice, and encouragement. I have learned from them how to develop, evaluate, express, and defend my ideas. These skills are essential for my MPhil study and my future academic career.

The System and Media Lab (SyMLab) with which I have been affiliated has been an excellent source of friendships as well as useful advice and collaboration from brilliant minds. In particular, I appreciate guidance and advice from Dr. Dimitris Chatzopoulos, Dr. Tristan Braud, and Dr. Lik Hang Lee. Special thanks to my fellow labmates in SyMLab: Ehsan Ul Haq, Reza Hadi Mogavi, Kirill A. Shatilov, Abhishek Kumar, and Ahmad Alhilal. We have finished several deadlines as a team. In daily life, we have been good friends. My graduate study in HKUST would not be so colorful without them.

I appreciate my parents and my family for their endless support, belief, and encouragement for me. I also would like to thank all of my close friends for always supporting and providing appropriate care with me.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title Page</td>
<td>i</td>
</tr>
<tr>
<td>Authorization Page</td>
<td>ii</td>
</tr>
<tr>
<td>Signature Page</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>viii</td>
</tr>
<tr>
<td>Abstract</td>
<td>ix</td>
</tr>
<tr>
<td>Chapter 1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Highlights of This Thesis.</td>
<td>2</td>
</tr>
<tr>
<td>Chapter 2 Related Work</td>
<td>5</td>
</tr>
<tr>
<td>2.1 User Engagement</td>
<td>5</td>
</tr>
<tr>
<td>2.2 Human Mobility and Revisitation Patterns</td>
<td>6</td>
</tr>
<tr>
<td>2.3 Growth and Evolution of Social Networks</td>
<td>7</td>
</tr>
<tr>
<td>2.4 Urban Computing and LBSNs</td>
<td>8</td>
</tr>
<tr>
<td>Chapter 3 Analytical Framework</td>
<td>9</td>
</tr>
<tr>
<td>3.1 Datasets</td>
<td>9</td>
</tr>
<tr>
<td>3.2 Characterization of User Types</td>
<td>10</td>
</tr>
<tr>
<td>3.3 Churn Prediction Problem</td>
<td>12</td>
</tr>
<tr>
<td>Chapter 4 Geographical Engagement Over User Lifecycle</td>
<td>14</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

3.1 Overview and a category hierarchy of location-based social networks. 10
3.2 CDF of time gaps between sequences of reviews of long-term producers. 12
4.1 Geographical engagement patterns. (a) The average radius of a user at each life-stage from an entire location history over her lifecycle. Users’ final $r_q(T)$ can be grouped into 4 distinct groups and the average radius of users converges from the initial stages of their lives. (b) The average radius of a user for her first $x$ reviews. It converges after the first few reviews. (c) The average moving distance of a user at each life-stage. Users’ moving distances are stable in most of their life-stages. Standard-error intervals are depicted but very small. (Same trends of (a), (b), and (c) in Yelp hold for (e), (f), and (g) in Foursquare, respectively) 15
4.2 The average probability of a user reviewing the vicinity of previously traveled locations in her immediate life-stage. Users keep exploring diverse venues in each life-stage over their lifecycle. 17
4.3 The average probability of revisiting the vicinity of all previously traveled locations of users for reviewing. Regardless of various threshold values of determining the distance of the vicinity ($d = 200, 400, 800, \text{and} 1200 \text{m}$), the probability converges. This result shows that users keep exploring geographically diverse venues with chances of at least 30–60% in Yelp and 50–70% in Foursquare over their lifecycle. 18
5.1 Churning probability according to the average radius and moving distance. 20
5.2 Revisiting probability according to $P_{\text{prev}}$ and $P_{\text{total}}$. 20
5.3 Average number of unique categories in each window. 22
5.4 The number of accumulated reviews on a venue in each window. 22
5.5 Churning probability according to the number of friends. 24
5.6 Churning probability according to the proportion of churned friends. 24
5.7 Linguistic aspects in Yelp. 25
5.8 Linguistic aspects in Foursquare. 25
6.1 Parameter sensitivity study on the performance of Stacked LSTMs. 33
LIST OF TABLES

3.1 Descriptive statistics for datasets. 11

6.1 List of Proposed Features of a User 29

6.2 Results of predicting whether a user will leave the community in the future. The prediction performances of LR models using each set of the proposed features and LSTM model using all features are presented for both Yelp and Foursquare. -Temporal represents LR model using all features except temporal feature, and the same holds for the rest of the features. 31

6.3 Feature Importance: $\chi^2$ Statistics 36
GEOGRAPHICAL ENGAGEMENT AND CHURN PREDICTION IN LOCATION-BASED SOCIAL NETWORKS

by

YOUNG DAE KWON

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology

ABSTRACT

As Location-Based Social Networks (LBSNs) have become widely used by users, understanding user engagement and predicting user churn are essential to the maintainability of the services. In this thesis, we conduct a quantitative analysis to understand user engagement patterns exhibited both offline and online in LBSNs. We employ two large-scale datasets which consist of 1.3 million and 62 million users with 5.3 million reviews and 19 million tips in Yelp and Foursquare, respectively. We discover that users keep traveling to diverse locations where they have not reviewed before, which is in contrast to “human life” analogy in real life, an initial exploration followed by exploitation of existing preferences. Interestingly, we find users who eventually leave the community show distinct engagement patterns even with their first ten reviews in various facets, e.g., geographical, venue-specific, linguistic, and social aspects. Based on these observations, we construct predictive models to detect potential churners. We then demonstrate the effectiveness of our proposed features in the churn prediction. Our findings of geographical exploration and online interactions of users enhance our understanding of human mobility based on reviews, and provide important implications for venue recommendations and churn prediction.
CHAPTER 1

INTRODUCTION

Location-Based Social Networks (LBSNs), with the most popular being Yelp and Foursquare, have become a vital part of our society due to their assistance on users’ needs. For example, foreign tourists in San Francisco can easily find highly-reputed restaurants even if it is their first time in the city. In addition, apart from assisting ordinary users every day, LBSNs also provide essential information, in the form of datasets and APIs, for researchers. Popular topics in LBSNs are: Point-of-Interest (POI) recommendation [93, 8, 89, 9], user mobility [15], privacy [80, 14, 86, 85], modeling users’ behaviors [19, 79, 53, 13], and urban computing [18].

There exist two types of LBSN users, the ones that produce information (e.g., by writing a review regarding a restaurant) and the ones that consume information (e.g., by reading reviews of the restaurants in one area). The producer-type users create User-Generated Content (UGC) that can affect the quality of experience of the consumer-type users [98] and the sustainability of the services by voluntarily sharing their location-related stories in LBSNs [59]. Considering that LBSNs heavily rely on UGC and users can stop contributing at any time, it is important for LBSN platforms to attract new users and keep the existing ones [39]. Hence, understanding the user engagement (i.e., the desire to use an application longer and repeatedly [47]) and predicting the user churn (i.e., the loss of a user from a service) is essential to the maintainability of the services [23, 99, 87].

The limitations of the existing studies can be categorised into two groups: (1) While there exist many studies investigating user engagement patterns in online settings [5, 54, 65, 21, 77, 31, 96], it is unclear how users explore and engage with LBSN services that can capture the offline and online experiences of the users. Do user engagement patterns in LBSNs present their own distinct patterns, or coincide with the human life course [25, 46]? That is, a person goes exploring in an “adolescent” phase and then becomes more stable by “settling down” later on. (2) Existing studies on churn prediction either focus on one user type such as newcomers [90, 23, 87], or one indicative feature set such as temporal [69, 81, 99],
linguistic [65, 77, 4], or social feature sets [68, 22]. Hence, the effect of each feature and combinations of different features are not yet well understood. In particular, the churning of users who are significantly more active (i.e., post more reviews) than average users has received less attention from the research community. Thus, in this thesis, we focus on highly active producer-type users while defining the scope of the user engagement to the user behavior of writing a review on the platform for further analysis.

To overcome the limitations of the first class, we conduct a quantitative study to understand user engagement patterns in LBSNs better. For that, we employ two large-scale datasets of LBSNs: Yelp and Foursquare. To address the second class of limitations, we first characterize user types according to their contribution levels to the services and then we further analyze the distinct engagement patterns that producer-type users with significant contributions manifest themselves in various aspects in the datasets. To this end, we use the LBSN datasets to answer the following research questions:

\[ \mathcal{RQ}_1 \] How do highly active producer-type users engage in the services of LBSNs in terms of geographical exploration?

\[ \mathcal{RQ}_2 \] How do engagement patterns of highly active producer-type users manifest themselves in various aspects?

\[ \mathcal{RQ}_3 \] To what extent can we predict the churning of users with significant contributions within a given period of time?

### 1.1 Highlights of This Thesis.

We present a large-scale quantitative exploration of patterns of user engagement in LBSNs. We employ a dataset from Yelp to analyze user engagement. Next we include a dataset drawn from Foursquare for making our results generalizable to LBSNs based on user reviews. In Chapter 3.1 we describe the details of these two datasets. In Chapter 3 we present an analytical framework used throughout this thesis. The focus of this study is on user engagement over the whole lifespan of users and on the precise prediction of churning users. Thus, we further characterize users who produce UGC into two long-term producers and ordinary producers to analyze their engagement according to the level of
contributions to the sites. We refer to those who made at least 50 reviews as *long-term producers* so that we can study users’ engagement for sufficiently long periods of their lifespans. Then, the remaining users, *i.e.*, the ones who wrote less than 50 reviews, are labeled as *ordinary producers*. This characterization facilitates our analyses on engagement, and churn prediction of long-term producers who write significantly more reviews (write 18 times more reviews in Yelp and 20 times more in Foursquare as it can be derived from Table 3.1) than ordinary producers in LBSNs. We then describe the churn prediction problem in detail. To answer RQ\(_1\), we examine how long-term producers engage with LBSNs in terms of geographical change. After that, we explore the differences in diverse aspects of long-term producers to answer RQ\(_2\). Lastly, we formulate a prediction task to answer RQ\(_3\).

To answer RQ\(_1\), we start with the assumption that users in LBSNs would be more likely to explore geographically and then become less adventurous with age. To understand user engagement patterns for sufficiently long periods of their lifespans, we employ long-term producers. We first find that users’ average radii and moving distances converge in a short time and are stable over their *lifecycle*, as defined in [21]. We then discover that users, in contrast to our initial assumption, continuously seek out different venues in new locations. We show that users return to the vicinity of previously reviewed venues from 10\% to 40\%, which means users tend to visit different venues with chances of 60–90\%. These results can give insights for site maintainers to offer personalized venue recommendations by considering users’ average radius and moving distances as well as their geographical engagement patterns.

For answering RQ\(_2\) and establishing principles to be used in prediction tasks, we examine the behavioral differences between churners and stayers among long-term producers from four aspects: (1) geographic, (2) venue-specific, (3) social, and (4) linguistic. Interestingly, behavioral differences between long-term producers who churn or stay are significant with their first 10 reviews. We find that long-term producers who stay consistently travel more to different locations and try more diverse categories of venues than those who churn. Besides, we discover that churned friends have more influence on long-term producers than on ordinary producers. Churning rates of long-term producers change more than twice than those of ordinary producers as the proportion of churning friends increases.
To answer RQ3, we formulate a churn prediction task to distinguish churning users from staying users. It is crucial to identify potential churners from the group of long-term producers before they decide to leave the community because this group produces about 40% of reviews on Yelp and 20% of tips on Foursquare according to the analyzed datasets. We demonstrate that the insights established in this thesis enable us to predict whether a user will stop contributing to the community in the future.

Our model based on Logistic Regression (LR) using all derived features achieves 0.768 AUC\(^1\), which outperforms all baseline models by up to 56.4% (43.3%) in AUC in Yelp (in Foursquare). After that, we explore to what extent we can further improve the performance of predicting churners by adopting a deep learning approach. Our best model achieves even higher performance of 0.882 AUC in Yelp and 0.799 AUC in Foursquare.

In summary, the main contributions of this thesis include:

1. We show that users constantly wander around diverse offline places, contrasting to the human life course assumption.
2. We find that the behavioral differences between churners and stayers are significant and that various factors show these differences with users’ first 10 reviews.
3. We demonstrate the effectiveness of our observations by significantly improving the performance over all of the baseline LR models on the churn prediction task. Based on our proposed features, we employ a deep learning model and achieve even higher performance in predicting potential churners in LBSNs.

The rest of the dissertation is organized as follows. Chapter 2 summarizes the related work, and Chapter 3 presents the designed analytical framework. In Chapter 4, we study users’ geographical engagement. After that, in Chapter 5, we examine the behavioral differences between churners and stayers. In Chapter 6, we predict churning users. After that, Chapter 7 discusses the limitations and the potential implications of this thesis. Finally, we conclude the thesis in Chapter 8.

\[1\] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
CHAPTER 2

RELATED WORK

We review the literature on user engagement, human mobility and revisitation patterns, and the growth and evolution of social networks as well as urban computing and LBSNs.

2.1 User Engagement

Researchers have investigated user engagement in various online settings such as mobile applications (apps) [54, 87], online communities [21, 27] as well as patterns [41, 62, 63] and motivations [71, 48] of user participation. In the literature, user engagement is formally defined as “the quality of the user experience that emphasizes the positive aspects of interacting with an online application and, in particular, the desire to use that application longer and repeatedly [47].” In addition, much work has focused on indicative features such as temporal [69, 81, 99], linguistic [65, 21, 77, 4], and social effects [68, 22] to improve user engagement or prevent users from churning. For instance, Danescu-Niculescu-Mizil et al. examined user engagement patterns from the language aspect and confirmed that users’ language usage becomes more inflexible with a community over time [21]. Amiri et al. used linguistic features from the tweets to predict the churning tweets for a specific brand of one telecommunication company among several others [4]. Instead of user churning from a single social network. Rashid et al. highlighted the similarity and acceptance of users with the rest of the community group are key factors for user participation on the platform [71]. Besides, Yang et al. used the network properties such as user’s network density and size with the user’s daily activities to predict churning users [87]. Lin et al. showed the primary intent of joining as the reason for multiple lives and proposed the method to predict the number of lives of the users [54]. Also, Mathur et al. modeled user engagement using contextual factors derived from smartphone usage and its embedded sensors [60]. Although Yang et al. [91] incorporated geographical information to improve the performance of the prediction of churning migrants from an urban area, it is still
unknown how users’ location histories relate to their engagement patterns in LBSNs. Thus, in this thesis, we analyze user engagement patterns concerning geographical exploration using two large-scale datasets of LBSNs. We further study the geographical influences as well as venue-specific, social, and linguistic features on user engagement according to user types.

2.2 Human Mobility and Revisitation Patterns

Many prior works studied to reveal human mobility patterns [27, 35, 15, 94, 16, 57, 56, 55]. Specifically, Gonzalez et al. using mobile phone data found that human mobility displays significant regularity because they return to a few highly frequented locations such as home or work [27]. Cho et al. demonstrated that human movement patterns are periodic both spatially and temporally as well as highly influenced by social network ties when the movement is long-distance [15]. Moreover, Choi et al. conducted a field study to show the significant association between geographical exploration and a user’s information seeking behavior [16]. Lu et al. characterized the lifecycle of POIs and developed a framework to predict the life status of POIs in a given time slot [57]. Meanwhile, there are studies that investigate revisitation patterns in online [67, 2, 70, 36] and offline contexts [11]. Obendorf et al. showed that users’ navigation strategies on web pages differ dramatically and are largely affected by their habits and type of a site visited. The authors categorized users’ revisitation patterns three-fold according to heuristically defined time as follows: short-term (within an hour), medium-term (within a week), and long-term (longer than a week) re-visits [67]. Adar et al. conducted a large-scale analysis of Web interaction logs of about 612K users and characterized four fundamental revisitation patterns (e.g., Fast, Medium, Slow, and Hybrid groups) by clustering the revisitation curves they proposed [2]. This work has been extended to smartphone app usage and human mobility in urban spaces. Jones et al. found that the revisitation behaviors of smartphone users on a macro-level resembles those of web browsing on desktops [36]. In the contexts of urban areas, Cao et al. analyzed the physical revisitations of individuals and compared both similarities and differences of online and offline revisitation patterns [11]. On the other hand, our work focuses on the reviewing behavior of a user which reveals one aspect of human mobility and other various factors manifested by users in LBSNs. We examine whether users show revisitation
patterns to previously visited locations or show exploratory patterns to new locations while relating our findings to a practical application such as churn prediction. Note that we consider re-visiting a vicinity of previously visited locations instead of visiting the exact same locations again by writing a review since the data shows inconsistent patterns: users in Yelp are highly unlikely to write a review at the same venue again (0.04%), whereas users in Foursquare on average write two tips on the same location. To the best of our knowledge, this is the first work to investigate individual mobility concerning writing reviews on POIs.

2.3 Growth and Evolution of Social Networks

Studies on the growth and evolution of social networks are well-recognized research topics [82, 42, 7, 51, 33, 38]. Kossinets and Watts analyzed an evolving social network using email interactions among students over regular semesters. The authors found that although local connections between individuals evolve over time, the overall network structure remains stable [42]. Backstrom et al. found dense communities which have more closed triads grow less [7]. Leskovec et al. examined the microscopic process of the evolution of social networks [51]. In contrast to a global social network, an ego network is a personal social network consisting of two components: an ego and alters [6, 45]. The ego is a single user, while alters are directly connected neighbors to the user. Kikas et al. studied ego networks on Skype and observed that bursty peaks of contact additions tend to appear shortly after user account creation [40]. Aiello and Barbieri analyzed the temporal evolution of ego networks extracted from Flickr and Tumblr and found that users tend to build most of their ego networks in the early stages of their life [3]. Our work is relevant to those studies since users’ churning behaviors can be influenced by their groups/communities (i.e., social network properties) as well as their underlying social networks (i.e., ego networks) as shown in prior works [68, 22]. In this thesis, we further extend those previous works by combining various ubiquitous data sources and significantly improve the performance of churn prediction tasks over a model built on social features.
2.4 Urban Computing and LBSNs

A new term, urban computing, has gained a significant attention which is largely attributed to the huge volume of data generated in cities [19, 9, 20, 58, 75, 74, 84, 12, 78]. These data-driven studies utilize different data sources ranging from GPS and mobile phone data to social media activity data. For example, Cranshaw et al. employed a GPS location tracking app to infer users’ social relations from their physical locations [19]. Xu et al. utilized mobile phone and travel survey datasets to represent temporal modes in human trajectory as well as showed the correlation between popular temporal modes and user’s occupations [84]. In order to improve the recommendation performance, prior works on POI [9] and event recommendation [20, 58] use users’ activity data as well as their social and geographical information. In addition, common practices to collect data involves self-reporting through surveys or collecting data through some observational methods [32]. Tu et al. tackled the cold-start problem of the personalized location recommendation by learning user interest and location features from app usage data [78]. In the contexts of LBSNs, Chen et al. investigated user behaviors of cross-site linking according to their privacy concerns [14]. D’Silva et al. investigated influential factors that cause the failure of retail businesses and developed predictive models to foretell business survival using Foursquare check-ins and transport data [24]. Furthermore, Yang et al. [88] proposed a hypergraph embedding approach designed for LBSNs data and improved the performance of friendship and location prediction tasks. The hypergraph includes user-user edges (i.e., friendships) and user-time-POI-semantic hyperedges (i.e., check-ins). Xu et al. designed a deep learning pipeline for fine-grained Location Recognition and Linking and then showed the effectiveness of their framework on Twitter data [83]. However, no prominent work has studied the churn prediction problem in the contexts of LBSNs.
CHAPTER 3

ANALYTICAL FRAMEWORK

In this chapter, we first describe the two datasets that we used in this thesis (Chapter 3.1). We then characterize two types of users according to their contributions (Chapter 3.2). After that, we describe the churn prediction problem (Chapter 3.3).

3.1 Datasets

We employ data from Yelp as our primary dataset to analyze user engagement. Then we include data from Foursquare to make our results more generalizable to LBSNs. The Yelp dataset is publicly available\(^1\) and spans from July 2004 to December 2017 in four English-speaking countries. It is composed of over 1.3 million users, and around 5.3 million reviews of 175 thousands of businesses\(^2\). The Foursquare dataset, collected by Chen et al. [13], spans from October 2008 to February 2016 from around the world. It includes over 62 million users and 19 million tips from 13 million venues.

Figure 3.1 describes the key components of LBSNs: (1) users, (2) venues, (3) reviews, (4) check-ins, (5) location history, and (6) category hierarchy. Each user has basic profile information such as name, ID, friend list, and profile photos. In addition, when users visit a venue (e.g., a shopping mall), they can leave a review and a rating for the venue. If users mark the venue, it is known as a check-in to the venue. For example, as shown in Figure 3.1a, \(u_1\) writes reviews of two venues which are \(l_1\) and \(l_2\). For each review, the location can be extracted from the venue that the review has been written about. We can construct the whole location history of users using their reviews.

In contrast to other works [18, 93] which use users’ check-ins to build their location histories, we use reviews to extract location histories of users for the following two reasons:

---

\(^1\)https://www.yelp.com/dataset

\(^2\)A review and business are called a tip in Yelp and venue in Foursquare. Hereafter we use these terms interchangeably. We restrict posts to reviews and use posts to refer to reviews.
Figure 3.1: Overview and a category hierarchy of location-based social networks.

(1) users often check-in to venues without physically visiting the locations as Zhang et. al. found that nearly 75% of all Foursquare check-ins do not match with the real mobility of users [97]; (2) writing a review usually indicates that a user performed the relevant activities like shopping or dining at the specified venue [8]. In LBSNs, venues are grouped into pre-defined categories and Figure 3.1b depicts the category hierarchy in LBSNs. For example, the category “Restaurants” includes “Korean Restaurants”, “American Restaurants”, and so on. Although there are three or four layers of category hierarchies in LBSNs, we use two layers from the root since more than half of the venues do not contain the third layer’s categories.

3.2 Characterization of User Types

Considering that LBSNs largely rely on UGC, we focus on users who create UGC and define user engagement as their reviewing behaviors in our study. The focus of this study is on user engagement over the whole lifespan of users as well as on the precise prediction of churning users. Hence, we further categorize producer-type users into two groups according to their levels of contributions to the sites. First, we refer to those who made at least 50 reviews as long-term producers so that we can study users’ engagement for sufficiently long periods of their lifespans. Second, we refer to the remainder of the users who wrote less than 50 reviews as ordinary producers. We set the number of contributions as the threshold used in prior works [21, 77] to distinguish long-term producers and ordinary
Table 3.1: Descriptive statistics for datasets.

<table>
<thead>
<tr>
<th></th>
<th>Yelp</th>
<th>Foursquare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of total producers</td>
<td>132,291</td>
<td>1,209,210</td>
</tr>
<tr>
<td>Number of total reviews</td>
<td>1,558,344</td>
<td>7,650,575</td>
</tr>
<tr>
<td>Number of 1st layer categories</td>
<td>22</td>
<td>9</td>
</tr>
<tr>
<td>Number of 2nd layer categories</td>
<td>598</td>
<td>411</td>
</tr>
<tr>
<td>Number of long-term producers</td>
<td>4,730</td>
<td>15,403</td>
</tr>
<tr>
<td>Number of reviews of long-term producers</td>
<td>621,373</td>
<td>1,524,843</td>
</tr>
<tr>
<td>Average number of reviews</td>
<td>131.4</td>
<td>99.0</td>
</tr>
<tr>
<td>Average number of categories</td>
<td>41.2</td>
<td>38.6</td>
</tr>
</tbody>
</table>

producers so that our findings on LBSNs can be directly comparable with the previous works.

We choose long-term producers as the main focus of our work to study user engagement patterns over time, following [21, 77]. We identify 4,730 long-term producers in Yelp\(^3\) (15,403 in Foursquare\(^4\)) who first posted reviews before January 2014 (March 2012) and wrote at least 50 reviews up until December 2016 (February 2015) in order to give them enough time (2014–2016 in Yelp and 2012–2014 in Foursquare) to accumulate 50+ reviews. It takes 820 and 720 days to accumulate 50 reviews on average in Yelp and Foursquare, respectively. Long-term producers are good subjects to study user engagement since their histories of activities in the community provide enough information to exhibit certain patterns. Moreover, long-term producers write 40% (20%) of reviews written by users who first posted in the period of consideration in Yelp (in Foursquare). Table 3.1 describes our final datasets which are used in the rest of this thesis. Note that a huge influx of users started from the year 2010 in both datasets, making a majority of long-term producers (99% in Foursquare and 89% in Yelp) who first posted after January 2009.

\(^3\)Since the Yelp dataset is a business-centered dataset, the dataset contains partial information of users’ whole review histories. Hence, we include users who have more than 50% of their whole review histories in our analysis. Whereas in the Foursquare dataset, entire tips of each user are collected with the maximum limit of 500 tips. In the end, we only sift out 512 long-term producers who have more than 500 tips in Foursquare.

\(^4\)Since the Foursquare dataset was crawled from around the world, we include users whose more than 80% of reviews are written in English for further analysis in terms of language aspects. We apply the same process on the Yelp dataset.
3.3 Churn Prediction Problem

Churn indicates the rate of loss of customers from companies in areas such as telecommunications [64, 34] or credit card services [66] where customers make a contract with a company. Like other online communities [90, 39, 23, 1, 29], users in online communities and LBSNs are not limited to a subscription contract. Hence, if a user $v$ is inactive for a substantial period, then user $v$ is considered to be a churned user, or otherwise a stayer. However, defining an inactive period as churn criterion differs depending on the context of the applications. To set the appropriate churn criterion in this thesis, we calculate the time gap of inactivity between two consecutive reviews of each user $v$. Figure 3.2 shows that over 95% of time gaps of users never exceed one year. We further check the re-engagement patterns of users which can be segmented into multiple disjoint active periods as suggested in [54]. When we set the inactive period as one year, around 60% in Yelp and 80% in Foursquare of long-term producers have only one lifespan (i.e., one active period). Thus, in this thesis, we adopt one year as the inactive period and analyze user engagement from a perspective of a single life rather than from a perspective of multiple lives.

Problem Statement

We aim to relate our analysis to users’ future activity status since it is important to predict the users’ future status in advance. Thus we further categorize users into two groups based on whether they eventually abandon the service or not. We first define a date (1 year before
the last date of the datasets) as the Start-Of-the-Future (SOF), similarly to [77]. The SOF is January 2017 in Yelp and March 2015 in Foursquare. We then define *staying users* as those who post at least one review as of the SOF; we define *churning users* as those who stop writing reviews as of SOF. As a result of our characterization of users and taking into account churn criterion, we identify 3,081 staying users and 1,649 churning users in Yelp as well as 6,153 staying users and 9,250 churning users in Foursquare. The ratios of staying to churning users are 65.1% to 34.9% in Yelp and 39.9% to 60.1% in Foursquare. Note that these long-term producers, who are the main focus of our study, had written at least 50 reviews before the SOF.

**Notation**

Given an LBSN $G(V, E)$, vertices are users and edges represent relationships between users. Set $V$ denotes the users and set $E$ their relationships. Each directed edge $e_{ij} \in E$ represents that user $v_i \in V$ follows another user $v_j \in V$. We call user $v_i$ a “follower” of user $v_j$. If user $v_j$ also follows user $v_i$, those users are “friends” with each other and we connect them with a directed edge $e_{ji}$. For each review users make, it contains location and time information. Using this information, we can keep track of users’ location histories (*i.e.*, sequences of venues the users have visited) along with the actual time period. In the following analysis, we use the time period as either *window* $W$ or *stage* $S$. Let user $v$ make $T$ reviews. Then the entire indexed sequence $1, ..., T$ can be grouped together as non-overlapping consecutive windows $w_i$ where $i \in [1, T/\text{window size}]$ and stages $s_j$ where $j \in [1, T/\text{number of stages}]$. The locations that user $v$ has visited can be denoted as $L^w_v = [l_1, ..., l_n]$, where location $l_k$ consists of latitude and longitude.
CHAPTER 4

GEOGRAPHICAL ENGAGEMENT OVER USER LIFECYCLE

In this chapter, we study user engagement from the perspective of users’ lifecycles considering all the user’s reviews. As in [21], we use the life-stages of a user to indicate the percentage of reviews the user has already written out of the total number of the reviews the user will write during her entire lifespan in the community. For example, a life-stage of 0%, birth, represents the time a user wrote her first review and a life-stage of 100%, death, represents the time a user leaves the community. In order to understand user engagement patterns over users’ overall lifecycles, we employ long-term producers who have contributed more than 50 reviews. We investigate how users explore the real world in terms of their visited locations revealed by the history of their reviews.

User’s average radius is determined in the early stage of their lifecycle. We use users’ location histories to understand their geographical exploration. Given the entire location history, $L_v = [l_1, ..., l_T]$, of user $v$, which is ordered by time and contains the latitude and longitude of the visited locations, we can compute user $v$’s average radius $r_g(t)$ using her trajectory up to $t^{th}$ reviews.

$$r_g(v, t) = \frac{\sum_{i=1}^{t} |l_i - l_{CM}|}{t}$$ (4.1)

where $l_{CM} = \frac{\sum_{i=1}^{T} l_i}{T}$ is the center of mass of her trajectory. As studied in [27], users can be grouped into distinct groups according to their final $r_g(T)$. We group them into four so that each group contains approximately 25% of users within it (e.g., 25% of users in Yelp and Foursquare are classified into a group with $r_g(T)$ less than 6 km and 80 km, respectively). Figures 4.1a and 4.1d show that the average radius of a user rapidly converges to $r_g(T)$ from the beginning of the user’s lifespan. Moreover, a similar trend appears when we change the x-axis to the number of reviews from the life-stage as shown in Figures 4.1b and 4.1e. We show the average radii of users using their reviews are determined in the
Figure 4.1: Geographical engagement patterns. (a) The average radius of a user at each life-stage from an entire location history over her lifecycle. Users’ final $r_g(T)$ can be grouped into 4 distinct groups and the average radius of users converges from the initial stages of their lives. (b) The average radius of a user for her first $x$ reviews. It converges after the first few reviews. (c) The average moving distance of a user at each life-stage. Users’ moving distances are stable in most of their life-stages. Standard-error intervals are depicted but very small. (Same trends of (a), (b), and (c) in Yelp hold for (e), (f), and (g) in Foursquare, respectively)

User’s moving distance is constant over the user’s lifecycle. Next, we examine the average moving distance of user $v$ at each life-stage of her entire lifecycle. Given user $v$’s location history at each life-stage $s$, $L_v^s = [l_1, ..., l_n]$, we calculate the average moving distance at each life-stage $s$ as follows:

$$\text{Average moving distance} = \frac{\sum_{i=1}^{n-1} |l_i - l_{i-1}|}{|L_v^s|}$$  (4.2)

Figures 4.1c and 4.1f show the average moving distances of users according to their $r_g(T)$. The average moving distance, similar to the average radius, is determined in the early stages of the users’ lifecycles. Besides, it remains consistent over the users’ entire lifetime,
meaning the users keep moving around geographically.

**Will users settle down or keep exploring geographically diverse venues for reviewing?** We examine whether user engagement patterns are analogous to the course of a human life [25, 46], i.e., a person explores in a “adolescent” phase and then stabilizes by “settling down”. To validate this question in terms of users’ geographical locations, we start by investigating the human movement by measuring how likely user \( v \) is to travel close to the venues that she has written reviews about before. We analyze how often users return to previously visited venues and in the vicinity of them by writing reviews. Intuitively, we expect that users tend to explore diverse venues in geographically different locations in their earlier life-stages and then tend to explore less while sticking to their preferred places. More specifically, the likelihood of reviewing geographically different locations is quite high in the beginning stages of the users’ lives, and then it decreases as users visit more and more places to review over their lifetimes in LBSNs. However, we will show below that this is not the case.

To quantify the propensity of users to explore diverse locations, we first define a counting function \( f \) to represent the number of occurrences of a user \( v \) reviewing the vicinity of the visited venues within radius \( d \) as follows:

\[
f(L^s_{i,v}, L^s_{j,v}) = |\{l_s \in L^s_{i,v} : \text{distance}(l_s, l_r) \leq d \land l_r \in L^s_{j,v}\}|
\]

(4.3)

where \( s_i \) is the current life-stage and \( s_j \) the previous life-stage. We then compute \( P_{\text{prev}}(L^s_{i,v}) \) which measures the probability that a user \( v \) in the current life-stage \( s_i \) revisits the vicinity of the venues for reviewing within radius \( d \) explored in the immediate life-stage \( s_{i-1} \) as follows:

\[
P_{\text{prev}}(L^s_{i,v}) = \frac{f(L^s_{i,v}, L^{s_{i-1}}_{i,v})}{|L^s_{i,v}|}
\]

(4.4)

\( P_{\text{prev}}(L^s_{i,v}) \) represents how likely the user \( v \) is willing to travel to various geographical locations in each stage of her life. For instance, \( P_{\text{prev}}(L^s_{i,v}) \) of 0% represents user \( v \) always writes reviews to different neighborhoods in her current life-stage \( s_i \) and \( P_{\text{prev}}(L^s_{i,v}) \) of 100% represents user \( v \) always return back to the vicinity of venues toured in her immediate life-stage \( s_{i-1} \). We analyze users’ geographical exploration using various threshold values of \( d \) defining the distance of the vicinity and observe similar trends. As shown in Figure 4.2, users in Yelp and Foursquare return to the vicinity of previously reviewed venues from 10%
Figure 4.2: The average probability of a user reviewing the vicinity of previously traveled locations in her immediate life-stage. Users keep exploring diverse venues in each life-stage over their lifecycle.

to 40% of the time. This result indicates that there is a 60–90% chance that users consistently post on different locations that they have not yet explored. Note that a radius of 400 meters is established as a standard for defining the size of a neighborhood which shares a similar functionality in the urban planning research community [95, 61].

We have observed that users are more likely to visit distinct locations when we only consider venues reviewed in their immediate life-stage $s_{i-1}$. However, people may return to the venues that they have reviewed in any life-stage $s_j$ where $j \in [1, i - 1]$. Hence, we calculate $P_{\text{total}}(L_{sv}^i)$ which measures the probability of the users in life-stage $s_i$ revisiting the vicinity of all of the traveled venues for reviewing until the immediate life-stage $s_{i-1}$ as follows:

$$P_{\text{total}}(L_{sv}^i) = \frac{\sum_{j \in [1, i-1]} f(L_{sv}^{s_i}, L_{sv}^{s_j})}{|L_{sv}^{s_{i-1}}|}$$  \hspace{1cm} (4.5)

The more users accumulate reviews, the easier it is for them to write reviews on venues which are in the vicinity of previously visited venues. Surprisingly, as in Figure 4.3 the probability of visiting different venues converges to 40–70% in Yelp and 30–50% in Foursquare. In other words, there exist 30–60% (50–70%) of chances that users in Yelp (in Foursquare) keep exploring geographically different neighborhoods depending on the threshold distances $d$ defining the vicinity. This result indicates that there is still a high chance that people travel to geographically distinct places even when we take into account
Figure 4.3: The average probability of revisiting the vicinity of all previously traveled locations of users for reviewing. Regardless of various threshold values of determining the distance of the vicinity (d = 200, 400, 800, and 1200 m), the probability converges. This result shows that users keep exploring geographically diverse venues with chances of at least 30–60% in Yelp and 50–70% in Foursquare over their lifecycle.

all of the venues that they have posted reviews on so far. As a result, we validate users’ geographical exploration throughout their lifecycle.

**Chapter summary.** The average radii and the average moving distances of users are settled soon after they start writing reviews on the site. However, users keep wandering around geographically diverse neighborhoods as they contribute more and more reviews in the site.
CHAPTER 5

ENGAGEMENT OF CHURNING AND STAYING USERS

After pointing out that users keep exploring diverse locations over their lifecycle in LBSNs, we now turn our attention to differences in user engagement between churners and stayers so that we can derive relevant feature sets to use in prediction tasks. In this chapter, to provide a holistic viewpoint to consider diverse aspects of ubiquitous data of LBSNs, we propose to investigate features related to location contexts such as geographical and venue-specific factors which LBSNs originally provide as well as social and linguistic factors as studied in previous works. Hence, we quantitatively study how churning behaviors manifest themselves among long-term producers using the following aspects: (1) geographic, (2) venue-specific, (3) social, and (4) linguistic aspects. To extract features for the prediction tasks, we take the initial 50 reviews rather than having $x\%$ of reviews because it is hard to know users’ entire lifespan and what percent of their life has passed before their departure. Hence, in this chapter, we conduct our analyses using the initial 50 reviews of producers. After that, in Chapter 6, we show the effectiveness of the derived features in churn prediction in which we attempt to detect churning users early in their lifecycle using the initial $x$ reviews.

5.1 Geographic Aspects

We delve into differences between churning and staying users manifested by geographical aspects. We choose four geographical features with which we examined the user engagement over the lifecycle in Chapter 4.

**Average radius.** We use the average radius to investigate how this feature is related to the churning rates of long-term producers. Figures 5.1a and 5.1b show the decreasing trend of churning rates as the average radius increases. Each point in Figure 5.1a and 5.1b corresponds to the $r_g(T)$ values used in Chapter 4. In Yelp, the churning rate is 50% at the
smallest $r_g(T)$ of 6 km. It becomes significantly reduced to 30% at $r_g(T)$ of 10 km and then is consistent onward. On the other hand, in Foursquare, the churning rate is 70% at the smallest $r_g(T)$ of 80 km. Then it continually reduces to 50% as the average radius increases.

**Average moving distance.** Similar to the average radius, the probability of churning rates is reduced according to the average moving distance as in Figures 5.1c and 5.1d. In Yelp, the trend of the churning probability of the average moving distance is shown to be almost identical to that of the average radius. In Foursquare, the churning rate declines from 65% to 50% as the average moving distance increases.

**Revisiting probability in an immediate window.** To further examine the churning behavior of users, we adopt the revisiting probability to the vicinity of venues in an immediate window based on written reviews of those users. Figures 5.2a and 5.2b show the average probability of writing reviews to the vicinity of venues that users visited in an immediate window. Staying users in Yelp are 4–6% and those in Foursquare are 2–3% less likely to write reviews than churning users from a neighborhood that they have already reviewed. This result indicates that both stayers and churners do not tend to return to the neighborhood of previously reviewed venues to write reviews again. Besides, stayers are relatively more likely to review new venues than churners.
Revisiting probability in all previous windows. Figures 5.2c and 5.2d shows the average probability of writing reviews to the vicinity of venues that users visited in all previous windows. Staying users in Yelp and Foursquare are around 4–6% on average less likely to write reviews than churning users from a neighborhood that they have already reviewed. This result further confirms two findings from the analysis of revisiting probability in an immediate window on human mobility based on reviewing behaviors of producer-type users.

5.2 Venue-specific Aspects

We employ several venue properties to study how user engagement relates to venues. Thus we use venue categories and the number of accumulated reviews written on a venue when a user wrote her review on it.

Venue categories. We employ second-layer categories (see Chapter 3.1 for detail) to analyze how venue properties affect user engagement. Figures 5.3a and 5.3b show the average number of unique second-layer categories in each window. Stayers are more likely to visit diverse categories of venues than churners. In addition, the same result also holds for other metrics such as entropy and Gini index. Both metrics are based on the probability of categories \( C \) in each window \( w_i \). The probability of a category \( c \) in a window \( w_i \) is computed as

\[
p_c = \frac{1}{|w_i|} \sum_{k \in w_i} I(C_k = c).
\]

Then the entropy and Gini-index in each window \( w_i \) are defined as Eq. 5.1 and Eq. 5.2 respectively.

\[
\text{Entropy} = - \sum_c p_c \log_2(p_c) \quad (5.1)
\]

\[
\text{Gini-index} = 1 - \sum_c p_c^2 \quad (5.2)
\]

In sum, users write reviews on categories more evenly in a window. In addition, churners focus on fewer categories.

Accumulated reviews on venues. We further investigate the venue properties using the number of reviews on a venue when a user visits it. We study whether the number of accumulated reviews on a venue affects users’ decisions to visit the venue. As shown in Figures 5.4a and 5.4b, we find that the average number of accumulated reviews on a venue increases for both churning and staying users. Moreover, churners write reviews on
venues with a fewer number of reviews accumulated than the venues on which stayers write reviews. Note that the y-axis of Figures 5.4a and 5.4b are normalized using a log with a base of 2, whose values are used as an input to classifiers in Chapter 6.

5.3 Social Aspects

Using social network properties of LBSNs, we examine the social influence on users’ churning behaviors. We compare the churning probability of long-term and all users to identify how much social influence affect churning behaviors of long-term producers who are the main focus of this thesis. Note that the employed data contains a static social network at the end of the inspection period. Hence, we do not have social network information of users when they first start writing reviews. However, based on the finding of the prior work that the ratio of added nodes to one’s social network is considerably
smaller after 20 days compared to the final ego-network size (Only nodes who have created links for at least 6 months are considered) [3], we could know that long-term producers’ social networks may stabilize over time. Since it takes on average 820 days in Yelp and 720 days in Foursquare to accumulate 50 reviews, we assume that social networks of long-term producers already stabilized when they write around 50 reviews. Based on this assumption, we analyze the social aspects of user engagement and then conduct an experiment only on 50 reviews in Chapter 6.

Degree. We first examine the relationship between churning probability and the number of friends that a user has. Figures 5.5a and 5.5b show decreasing trends in churning rates as the number of friends increases. This tendency matches with our common sense that users with many friends in the service are less likely to leave. Note that we use the number of friends as a degree because the datasets do not provide follower relations but friend relations. In sum, we show that churning rates of all users are consistently higher than those of long-term producers. Furthermore, we confirm that the churning probability declines for all and long-term producers in proportion to the degree of the users.

Proportion of churned friends. Figures 5.6a and 5.6b show the churning rates of all and long-term producers according to the proportion of their churned friends, respectively. We find there are more churning users as they have more and more churned friends, as studied in [68]. In our study, we take all users as a reference point to examine the churning behaviors of long-term producers. The churning rates are very high of 80–100% for all users and 30–70% for long-term producers at the X-axis of 0%. These high churning rates are reasonable since the X-axis of 0% occurs when the user has no friends. Furthermore, we discover that long-term producers are more sensitive to their friends’ churning. As in Figure 5.6a, churning rates of long-term producers increase significantly from 5% to 60% as the proportion of churned friends is raised from 20% to 80% in Yelp, which is almost twice as much change in the churning rates of all users. Also, Figure 5.6b shows a more significant change in churn rates of long-term producers in Foursquare. Thus, while the churning rates of all users are always higher than those of long-term producers, long-term producers are more sensitive to their friends’ departure from the community.
5.4 Linguistic Aspects

We finally investigate how linguistic aspects affect users’ churning behaviors using review text. We herein adopt language patterns that users use as a proxy to look into users’ online interactions and engagement patterns in the community. We first take review lengths. Then, as studied in [17, 73], we take the frequency of pronouns to study the differences between churners and stayers.

**Review length.** We find that long-term producers in both Yelp and Foursquare write longer reviews while they accumulate more and more reviews on the site, as shown in Figures 5.7a and 5.8a. Moreover, in Yelp, the review length of churning users is significantly longer than that of staying users. In Foursquare, churning users write longer reviews in their first 10 reviews. However, the difference in review length between churning and staying users disappears after they accumulate more than 10 reviews. It is noteworthy...
that long-term producers in Yelp write much longer texts than the users in Foursquare. These differences between Yelp and Foursquare in terms of review length could be derived from the characteristics of each LBSN. Specifically, Yelp encourages long and detailed reviews on venues by displaying an elaborate and exemplary review to users. On the other hand, Foursquare promotes concise and brief tips on venues by asking a simple question (e.g., What’s good here?) with the limit of word count on tips. Although we identify the behavioral differences between stayers and churners in both Yelp and Foursquare, the result is hard to be generalizable across the different platforms of LBSNs. As discussed above, it is because the length of the reviews can be highly different in each LBSN depending on its interface design and particular limitation on the word count of reviews.

**Frequency of pronouns.** Prior works [21, 77] studying linguistic features in online communities suggested that the decreasing frequency of first-person singular pronouns (e.g., I, Me) can indicate the users’ increasing identification with the community. In our study, however, Figures 5.7b and 5.8b show the opposite trend, i.e. the frequency of first-person singular pronouns manifest increasing patterns as users post more reviews on...
If we follow the interpretation that the low frequency of first-person singular pronouns is associated with a higher level of affiliation with the community, it is hard to explain the increasing patterns of pronoun usage in LBSNs because it means newcomers who will churn have the highest affiliation with the community. Thus, we present a new perspective based on the linguistic theory that can better explain the observed language patterns in LBSNs.

In the literature, the frequencies of pronouns used by a user can be associated with the focus of the user [17, 73, 37]. For example, if a user often uses first-person pronouns (e.g., I, Me, We, Us), this indicates that the user’s attention is on herself, friends, or family members within her group. On the other hand, the frequent usage of second and third-person pronouns (e.g., you, yours, they, theirs) can be associated with a user’s attention on others who are not necessarily within her group. Users in LBSNs are likely to use more first-person pronouns as they contribute reviews to the community (see Figures 5.7b and 5.8b). In contrast, Figures 5.7c and 5.8c show that the usage of second and third-person pronouns represents consistent patterns over users’ lifespans. This result may indicate that users focus more and more on people within their groups. Note that a small fraction of difference in pronoun frequencies can reveal meaningful behavioral differences in language patterns of users as in [77]. However, since the actual difference is quite small, linguistic feature sets extracted from frequencies of the pronouns are not as effective as other proposed features in the prediction task in Chapter 6 (see Table 6.3). Also, our analysis of the frequency of pronouns used by LBSN users is based on the English language. Hence, it is not clear how this result can be applied to other languages such as German, Chinese, or Korean.

1 Although Figures 5.7b and 5.8b represent both first-person singular and first-person plural pronouns, first-person singular pronouns show a similar trend of Figures 5.7b and 5.8b.
CHAPTER 6

PREDICTING CHURNING USERS

Having established engagement patterns from various aspects, we turn now to study to what extent we can predict churning users from staying users using the identified engagement patterns. Predicting churners among a group of long-term producers has practical value for creating and maintaining online communities since the LBSNs primarily rely on the type of users who actively contribute 40% of all reviews. Furthermore, given that almost 70% of customers in subscription services will not come back once they stop the subscription [43], early detection of long-term contributors who are likely to leave the community is crucial to service maintainers by enabling them to use many strategies to re-involve the users in the services before they abandon those services.

6.1 Experimental Setup

We formulate a prediction task of detecting churning users using an initial $k$ reviews. We then adopt the oversampling method to tackle the class imbalance of the datasets. Moreover, we vary the initial $k$ reviews of users to examine the performance changes according to various first $k$ reviews used for training classifiers.

Baseline

Prior work on user engagement found that a temporal feature, time gaps between reviews, is a powerful indicator to predict users’ churning [69] and has often been used for training classifiers in recent works for user engagement in online communities [21, 77]. We build a benchmark using this temporal feature as a reference point and compare it with our proposed features to show the effectiveness of our proposed features on the classification task.
Proposed Features

We propose features based on the observations that we have reported in the previous chapters. Table 6.1 describes four sets of features in detail. For all features except for social features, we use the aggregated values in each window with a size of 10 posts. Hence, we have five values for each feature. To take into account the temporal dynamics of features, we include the index of the window with the maximum and minimum value. Then we use each set of features to build the models for churn prediction task as follows.

(F_1) **Temporal feature:** Average time gaps between reviews.

(F_2) **Geographic feature:** Average radius, average moving distance, \( P_{\text{prev}} \), and \( P_{\text{total}} \).

(F_3) **Venue properties:** Unique category, entropy, gini-index, and average number of accumulated reviews on a venue.

(F_4) **Social feature:** Number of friends and churn rates of friends.

(F_5) **Linguistic feature:** Average frequency of first, second, and third-person pronouns as well as average number of words.

To further investigate the most prominent features based on our observations in Chapter 5 and Chapter 6 for learning models, we build models using the top-2 important features listed in Table 6.3. Also, we build models using the top-2 important features and two geographic features, one of the primary contributions in both Chapter 5 and Chapter 6. Finally, we construct the full model using all features and leave-one-out models from the full model.

(F_6) **Top2:** Average time gaps between reviews and average number of accumulated reviews on a venue.

(F_7) **Top2+Geo2:** Average time gaps between reviews, average number of accumulated reviews on a venue, average radius, and average moving distance.

(F_8) **All:** Combination of all features.

(F_9:15) **Leave-one-out:** Combination of all features without one feature set.
Table 6.1: List of Proposed Features of a User \( v \)

### Geographic Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Radius</td>
<td>Average distance of ( v ) up to ( t^{th} ) review from her center of mass ( l_{CM} ). ( i.e., \frac{1}{t} \sum_{i=1}^{t}</td>
</tr>
<tr>
<td>Moving Distance</td>
<td>Average distance that ( v ) moves in window ( w ). ( i.e., \frac{1}{</td>
</tr>
<tr>
<td>( P_{\text{prev}} )</td>
<td>Probability ( P_{\text{prev}}(L_v^{w_i}) ) that ( v ) visits venues within radius ( d ) of venues reviewed in an immediate window ( w_{i-1} ) in current window ( w_i), i.e., ( P_{\text{prev}}(L_v^{w_i}) = \frac{f(L_v^{w_i}, L_v^{w_{i-1}})}{</td>
</tr>
<tr>
<td>( P_{\text{total}} )</td>
<td>Probability ( P_{\text{total}}(L_v^{w_i}) ) of ( v ) in current window ( w_i ) reviewing the vicinity of all of the venues ( v ) has travelled for all windows ( w_j ) where ( j \in [1, i-1], i.e., P_{\text{total}}(L_v^{w_i}) = \frac{\Sigma_{j \in [1, i-1]} f(L_v^{w_i}, L_v^{w_j})}{</td>
</tr>
</tbody>
</table>

### Venue Properties

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique Category</td>
<td>Average number of unique second-layer categories.</td>
</tr>
<tr>
<td>Entropy</td>
<td>Category diversity based on the probability of categories ( C ) in each window ( w_i), ( i.e., -\Sigma c p_c \log_2(p_c). )</td>
</tr>
<tr>
<td>Gini-index</td>
<td>Category diversity based on the probability of categories ( C ) in each window ( w_i), ( i.e., 1 - \Sigma c p_c^2 ) where probability of a category ( c ) is computed as ( p_c = \frac{1}{</td>
</tr>
<tr>
<td># Accu. Reviews</td>
<td>Average number of accumulated reviews on a venue when ( v ) write a review on the venue.</td>
</tr>
</tbody>
</table>

### Social Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>Number of friends ( v ) has.</td>
</tr>
<tr>
<td>% Churned Friends</td>
<td>Percentage of churned friends of ( v ).</td>
</tr>
</tbody>
</table>

### Linguistic Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review Length</td>
<td>Average number of words in ( v )'s reviews.</td>
</tr>
<tr>
<td>1st person</td>
<td>Average frequency of first-person pronouns used by ( v ).</td>
</tr>
<tr>
<td>2nd person</td>
<td>Average frequency of second-person pronouns used by ( v ).</td>
</tr>
<tr>
<td>3rd person</td>
<td>Average frequency of third-person pronouns used by ( v ).</td>
</tr>
</tbody>
</table>

### Methods for Evaluation

We use Logistic Regression (LR) with L2-regularization as a classifier to predict churners. We adopt the LR model which can provide us with highly interpretable information on our derived features since the primary goal of our study is to identify and validate important features from the observations we made in Chapter 5 and 6. Thus, we train
LR for all proposed combinations of features (i.e., $F_1:F_{15}$). Furthermore, inspired by the recent advancements in the deep learning approach for sequential data [50] [87], we adopt the Long Short-Term Memory (LSTM) recurrent neural network (RNN) widely used for time-series analysis [28] [52]. After that, to explore to what extent we can further enhance the performance of the churn prediction task, we train LSTM using all features ($F_8$) which shows the best performance among all LR models. In experiments for LSTM models, we leverage Adam [49] as the optimizer and implement them with TensorFlow architecture. We set the batch size and learning rate to 32 and 0.001, respectively. Besides, we adopt the Glorot initialization [26] and early stopping [72] in the training process. The dropout probability [76] is set to 0.1 at the last LSTM layer. For both LR and LSTM models, we optimize the hyperparameters with the grid search strategy.

**Evaluation Protocols**

We define the task to predict whether a user will churn or stay after their 50th review. To distinguish churning producers from staying producers, we extract features based on users’ first $k$ reviews where $k = 10, 20, 30, 40, 50$. For social features, we conduct experiments only with the case of 50 reviews due to the limitation of static social network data. Since the proportions of churners and stayers are imbalanced, we balance the proportion of two classes by oversampling the minority class (i.e., churners) [10]. In order to overcome the potential bias in our sampled datasets and to obtain the generalizability of our results, we conduct the experiments over 20 randomly sampled datasets. We determine 90% of users as training/evaluation sets and the remaining 10% of users as a test set, respectively. Then we use the area under the receiver operating characteristic curve (AUC) to evaluate the performance of models. AUC is a widely used measure to assess the performances of classifiers in imbalanced data [10] [54].

### 6.2 Evaluation on LR

The performances for predicting whether a user will depart the community in the future are shown according to the number of first $k$ reviews used for training and testing the LR models (see Table 6.2). LR models trained with Top2 and Top2+Geo2 features show
Table 6.2: Results of predicting whether a user will leave the community in the future. The prediction performances of LR models using each set of the proposed features and LSTM model using all features are presented for both Yelp and Foursquare. -Temporal represents LR model using all features except temporal feature, and the same holds for the rest of the features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>k = 10</th>
<th>k = 20</th>
<th>k = 30</th>
<th>k = 40</th>
<th>k = 50</th>
<th>k = 10</th>
<th>k = 20</th>
<th>k = 30</th>
<th>k = 40</th>
<th>k = 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal</td>
<td>0.659</td>
<td>0.692</td>
<td>0.696</td>
<td>0.700</td>
<td>0.700</td>
<td>0.634</td>
<td>0.668</td>
<td>0.681</td>
<td>0.683</td>
<td>0.683</td>
</tr>
<tr>
<td>Geographic</td>
<td>0.589</td>
<td>0.605</td>
<td>0.604</td>
<td>0.609</td>
<td>0.609</td>
<td>0.577</td>
<td>0.594</td>
<td>0.604</td>
<td>0.606</td>
<td>0.605</td>
</tr>
<tr>
<td>Venue</td>
<td>0.616</td>
<td>0.622</td>
<td>0.640</td>
<td>0.659</td>
<td>0.661</td>
<td>0.566</td>
<td>0.572</td>
<td>0.587</td>
<td>0.598</td>
<td>0.606</td>
</tr>
<tr>
<td>Social</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.714</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.619</td>
</tr>
<tr>
<td>Linguistic</td>
<td>0.583</td>
<td>0.599</td>
<td>0.602</td>
<td>0.607</td>
<td>0.613</td>
<td>0.520</td>
<td>0.540</td>
<td>0.545</td>
<td>0.545</td>
<td>0.547</td>
</tr>
<tr>
<td>Top2</td>
<td>0.675</td>
<td>0.704</td>
<td>0.708</td>
<td>0.713</td>
<td>0.713</td>
<td>0.624</td>
<td>0.664</td>
<td>0.680</td>
<td>0.682</td>
<td>0.683</td>
</tr>
<tr>
<td>Top2+Geo2</td>
<td>0.681</td>
<td>0.710</td>
<td>0.713</td>
<td>0.722</td>
<td>0.723</td>
<td>0.630</td>
<td>0.668</td>
<td>0.686</td>
<td>0.688</td>
<td>0.689</td>
</tr>
<tr>
<td>All</td>
<td>0.687</td>
<td>0.715</td>
<td>0.720</td>
<td>0.729</td>
<td>0.768</td>
<td>0.633</td>
<td>0.671</td>
<td>0.689</td>
<td>0.692</td>
<td>0.711</td>
</tr>
<tr>
<td>–Temporal</td>
<td>0.643</td>
<td>0.659</td>
<td>0.669</td>
<td>0.684</td>
<td>0.736</td>
<td>0.594</td>
<td>0.615</td>
<td>0.628</td>
<td>0.634</td>
<td>0.661</td>
</tr>
<tr>
<td>–Geographic</td>
<td>0.680</td>
<td>0.706</td>
<td>0.716</td>
<td>0.722</td>
<td>0.767</td>
<td>0.624</td>
<td>0.667</td>
<td>0.683</td>
<td>0.685</td>
<td>0.708</td>
</tr>
<tr>
<td>–Venue</td>
<td>0.673</td>
<td>0.709</td>
<td>0.713</td>
<td>0.721</td>
<td>0.768</td>
<td>0.630</td>
<td>0.670</td>
<td>0.688</td>
<td>0.692</td>
<td>0.710</td>
</tr>
<tr>
<td>–Social</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.728</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.691</td>
</tr>
<tr>
<td>–Linguistic</td>
<td>0.680</td>
<td>0.712</td>
<td>0.717</td>
<td>0.723</td>
<td>0.763</td>
<td>0.634</td>
<td>0.670</td>
<td>0.686</td>
<td>0.689</td>
<td>0.710</td>
</tr>
<tr>
<td>–Top2</td>
<td>0.611</td>
<td>0.637</td>
<td>0.638</td>
<td>0.647</td>
<td>0.721</td>
<td>0.589</td>
<td>0.609</td>
<td>0.619</td>
<td>0.622</td>
<td>0.654</td>
</tr>
<tr>
<td>–(Top2+Geo2)</td>
<td>0.586</td>
<td>0.630</td>
<td>0.634</td>
<td>0.642</td>
<td>0.722</td>
<td>0.541</td>
<td>0.566</td>
<td>0.577</td>
<td>0.577</td>
<td>0.643</td>
</tr>
<tr>
<td>Stacked LSTMs</td>
<td>0.735</td>
<td>0.844</td>
<td>0.847</td>
<td>0.858</td>
<td>0.882</td>
<td>0.664</td>
<td>0.751</td>
<td>0.770</td>
<td>0.773</td>
<td>0.799</td>
</tr>
</tbody>
</table>
improvement over the benchmark model, which indicates that the number of accumulated reviews and two geographical features provide additional information onto the benchmark model trained with a strong temporal feature. In addition, our full LR model which utilizes all features outperforms all other models, by achieving 0.768 AUC in Yelp (0.711 AUC in Foursquare). Note that a random baseline will show 0.50 AUC. The full LR model also significantly improves the performance of the benchmark which uses strong indicator, temporal feature, by 9.7% (4.2%) in AUC and other baseline models by up to 56.4% (43.4%) in AUC in Yelp (in Foursquare). The differences between the benchmark and full LR model for both prediction tasks in Yelp and Foursquare are statistically significant according to the Wilcoxon signed-rank test ($p < 0.001$). This result validates the effectiveness of our suggested features in distinguishing churning users.

Furthermore, we conduct experiments with LR models using leave-one-out feature sets (i.e., rows of –Temporal, –Geographic, –Venue, –Social, –Linguistic, –Top2, and –(Top2+Geo2)) to scrutinize how performance decreases with cutting part of the component. When using the first $k = 10, 20, 30, 40$ reviews, the performance of all leave-one-out LR models decreases. However, when using the first $k = 50$ reviews, the addition of Geographic and Venue features to LR models does not lead to performance improvement. It seems that geographic and venue-specific features do not provide much information by including more reviews since those features are consistent over time. On the other hand, adding social features when using the first $k = 50$ reviews significantly improves the overall performance of LR models. This result indicates that using all possible features is essential for an improved performance as the full LR model performs the best.

### 6.3 Evaluation on Stacked LSTMs

Based on the result of LR models, we conduct further experiments to investigate to what extent we can improve the overall performance leveraging the recent advancement of a deep learning approach. For that, we utilize Stacked LSTM recurrent neural networks using all features to compare with the full LR model. Table 6.2 also lists the evaluation results of Stacked LSTM models according to the first $k = 10, 20, 30, 40, 50$ reviews with both Yelp and Foursquare datasets. We observe that Stacked LSTM outperforms the best
performing LR model over all cases. For the Yelp dataset, Stacked LSTM improves 6.9–18.0% over the best LR model. Also, Stack LSTM consistently achieves the best performance by improving 4.8–12.4% over the best LR model in the Foursquare dataset. In the end, the best Stacked LSTMs achieve a high AUC of 0.882 in Yelp and 0.799 in Foursquare.

Furthermore, we conduct experiments to investigate parameter sensitivity. Stacked LSTM involves several parameters (e.g., hidden state dimension, the number of stacked LSTM layers, batch size, dropout probability). To examine the robustness of the trained Stacked LSTM models, we investigate how the performance of Stacked LSTM in predicting churning users is affected by the different choices of parameters. Except for the tested parameter (i.e., hidden state dimension and the number of stacked LSTM layers), we set other parameters to the default values as specified in §6.1. Figure 6.1 shows the evaluation results of Stacked LSTMs by varying two parameters. First of all, we observe that the change in Stacked LSTMs’ performance is minimal when $k = 10$ with both parameters. In addition, we observe that the increase in the performance saturates as the hidden state
dimension reaches around 64, which demonstrates that the larger dimensionality does not always bring performance increase. On the other hand, we find that the number of LSTM layers have a relatively low impact on the performance of Stacked LSTMs. This result indicates that even with a single layer LSTM can achieve high performance in predicting producer-type users in LBSNs.

6.4 Performance Change Over Reviews

We further discuss the change in performance with respect to the number of initial $k$ reviews used for training the models. The performance of all models increases as we use more reviews for training the models as shown in Table 6.2. For example, a full LR model using 50 reviews displays an additional 11.7% (12.3%) improvement in AUC over the full LR model using 10 reviews in Yelp (in Foursquare). Similarly, the performance of Stacked LSTMs based on $k = 50$ posts improves by 20.1% (20.4%) in AUC over the model based on $k = 10$ posts in Yelp (in Foursquare). This result indicates that considering temporal dynamics by acquiring more reviews is as essential as having informative features. Note that the Stacked LSTM model using the first 10 reviews also achieves 0.735 AUC in Yelp (0.664 AUC in Foursquare). It represents that the information in the earlier stage of a user’s life provides enough predictive power to predict churning users accurately. This result is impressive since we can already make an accurate prediction of the future status of users from their first 10 posts.

6.5 Understanding Feature Importance

We finally investigate the feature importance of our proposed features. We calculate the $\chi^2$ (Chi-square) statistic to evaluate the discriminative power of our proposed features [92]. Table 6.3 shows the top 10 most important features with $\chi^2$ scores of Yelp and Foursquare datasets. Along with the temporal feature, features such as geographic and venue properties derived from offline context are top 5 important features. As we discussed the results of experiments with leave-one-out features, it may not be informative when we have full 50 reviews of producers. However, this result can indicate the impact of taking offline context
into account for the prediction task in more common cases where we only have the limited information of users (e.g., 10 reviews or less). Other features such as linguistic, social features are also vital for constructing the powerful predictive model since the models using the combination of all features perform the best.
<table>
<thead>
<tr>
<th>Rank</th>
<th>(\chi^2)</th>
<th>Feature Category</th>
<th>Feature Category</th>
<th>(\chi^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16625.17</td>
<td># Accu. Reviews</td>
<td>Venue</td>
<td>8301.76</td>
</tr>
<tr>
<td>2</td>
<td>5533.68</td>
<td>Time Gap</td>
<td>Temporal</td>
<td>7801.16</td>
</tr>
<tr>
<td>3</td>
<td>76.07</td>
<td>% Churned Friends</td>
<td>Social</td>
<td>439.84</td>
</tr>
<tr>
<td>4</td>
<td>34.92</td>
<td>Average Radius</td>
<td>Geographic</td>
<td>344.47</td>
</tr>
<tr>
<td>5</td>
<td>28.92</td>
<td>Moving Distance</td>
<td>Geographic</td>
<td>23.54</td>
</tr>
<tr>
<td>6</td>
<td>14.36</td>
<td>Degree</td>
<td>Social</td>
<td>9.46</td>
</tr>
<tr>
<td>7</td>
<td>12.36</td>
<td>(p_{\text{prev}})</td>
<td>Geographic</td>
<td>8.54</td>
</tr>
<tr>
<td>8</td>
<td>12.01</td>
<td>(p_{\text{total}})</td>
<td>Geographic</td>
<td>8.77</td>
</tr>
<tr>
<td>9</td>
<td>6.22</td>
<td>Review Length</td>
<td>Linguistic</td>
<td>8.11</td>
</tr>
<tr>
<td>10</td>
<td>1.95</td>
<td>Unique Category</td>
<td>Venue</td>
<td>4.64</td>
</tr>
</tbody>
</table>

Table 6.3: Feature Importance: \(\chi^2\) Statistics
CHAPTER 7

DISCUSSION

We use this chapter first to discuss some limitations to this thesis (Chapter 7.1). After that, we summarize our findings in users’ geographical exploration patterns and user engagement in four different aspects (Chapter 7.2). Note that here we present the correlation between user behaviors and examined features, not causal relationships. Finally, we discuss the potential applications of our findings (Chapter 7.3).

7.1 Limitations

There are limitations to this thesis due to the employed datasets as follows:

1. **Data availability.** The Yelp dataset is a business-centered dataset, which contains whole review histories of businesses but only some portion of those of users. Although we have tried our best to remain the highest coverage for analysis, many Yelp users are sifted out during the preprocessing process. Thus, there may exist some bias in the analysis due to data availability. However, in the Foursquare dataset, we capture the whole review histories of users. Moreover, our analysis shows consistent patterns for both Yelp and Foursquare users in the geographical, venue-specific, and social aspects. The bias can be largely mitigated.

2. **Studied users.** We focus on the behaviors of long-term producers who contributed at least 50 reviews so that we had sufficient history per user to observe her trajectory in the community as well as in the real world. However, this user type makes up 3.6% (1.3%) of the user base in Yelp (Foursquare) during the studied period, which limits our study to a small portion of users in LBSNs. Furthermore, our study excluded consumer-type users since they do not leave any logs to analyze in our study.
3. **Engagement Diversity.** LBSNs enable users to interact with the services in many ways. For example, users can search through the sites to find the next destination to dine, read reviews of local restaurants in unexplored areas to decide their visits to the venue, and interact with other users through up-voting their reviews or by following them. In our study, we could not consider diverse aspects of user engagement such as reading reviews and interacting with other users, since the employed datasets do not contain such information. However, by narrowing down the scope of the user engagement of producer-type users to the activity of writing a review, we present interesting findings in this thesis, which leads us to useful implications as we discuss in Chapter 7.

### 7.2 Summary of Findings

The following six points can summarize our quantitative study on LBSNs:

(\(P_1\)) The average radii and moving distances of users are determined within 5–10 reviews (Chapter 4).

(\(P_2\)) Users consistently write reviews on different locations at least 50% of all reviews for each life-stage (Chapter 4).

(\(P_3\)) Staying users are more likely to explore diverse locations than churning users (Chapter 5.1).

(\(P_4\)) Staying users are more likely to write reviews on venues of diverse categories and with more reviews accumulated (Chapter 5.2).

(\(P_5\)) The probability of churning increases for users with a higher percentage of churning friends (Chapter 5.3).

(\(P_6\)) Churning users, in Yelp, use less first-person pronouns and write longer reviews. In Foursquare, on the other hand, churning users use more first-person pronouns and reviews of approximately the same length as the staying users (Chapter 5.4).
7.3 Implications

Based on the findings on users’ geographical exploration patterns, we first confirm the previous studies on human mobility that human movement patterns are periodic and regularized [27, 15]. We also observe that users keep reviewing diverse categories and different locations of venues in contrast to the human life course theory [25, 46, 21], where a person explores in a “adolescent” phase and then stabilizes by “settling down”. Our finding is in accordance with the prior work on users’ community seeking behaviors in Reddit [77]. Furthermore, all of our discoveries, including various aspects of engagement patterns of users, have implications for site maintainers to increase user engagement in LBSNs. To increase the engagement levels of users, incentive mechanisms like gamification [30] can be employed to encourage engagement. For example, for those users who are at risk of departing, site maintainers can provide some incentives such as rewards and badges or can recommend different venues that users have never reviewed before to re-engage them.

Our findings can be utilized in the following ways:

1. Based on (P1), early recognition of user geographical exploration patterns enable the site maintainers to provide a more personalized user experience. Since the average radii and moving distances of users are determined within 5–10 reviews and are stable over their lifespan, one can recommend venues located within a user’s average radius and moving distance from her center of mass. For example, for those who have a small average radius and moving distance less than 6km, one can suggest nearby venues to the users. On the other hand, for those who have a large average radius and moving distance greater than 100km, one can even recommend venues located in another city.

2. Based on (P2) and (P4), since users tend to write reviews on unexplored geographical locations and diverse categories of venues, we can recommend a venue with a new category and location in unexplored neighborhoods that the user has not yet visited for reviewing. For example, if a user has some reviews on a category of “Mexican Restaurant” in one neighborhood, one can recommend a “Chinese Restaurant” located in another location to encourage the user to explore and to increase her engagement with the services.
3. Based on \((\mathcal{P}_4)\), staying users are more likely to write reviews on popular venues (i.e., the high number of accumulated reviews). It seems that users are more satisfied with popular venues. Hence, we can suggest popular venues to users for reviewing to increase their engagement on the services.

4. Based on \((\mathcal{P}_3), (\mathcal{P}_4), (\mathcal{P}_5),\) and \((\mathcal{P}_6)\), the powerful predictive model leveraging various data sources of geographical, venue-specific, social, linguistic aspects enables the site owners to detect users who are about to churn. After identifying those users who have a high probability of churning, one can employ gamification techniques such as badges and rewards to motivate them not to leave the service. For example, they could be awarded for additional reviews after a long period of inactivity.
CHAPTER 8

CONCLUSION

In this thesis, we studied the engagement patterns of producer-type users based on various aspects including geographical, linguistic, venue-specific, and social features. We performed a large-scale analysis of the representative LBSNs (i.e., Yelp and Foursquare). We initially characterized user types on the employed large-scale datasets to focus our analysis on long-term producers who contribute the most UGC to the community among all user types. After that, we examined how long-term producers behave geographically in the offline real world and engage with the online community of LBSNs. First, in contrast to the human life course assumption, we found that users exhibit exploring behaviors until the end of their life in LBSNs. For example, they consistently travel to different locations at least 60% of all reviews for each life-stage. Second, we found that churning users and staying users show different patterns in four aspects. To name a few, staying users are more likely to travel to unexplored neighborhoods for reviewing and write reviews on diverse venues with more accumulated reviews. Besides, from the social aspect, we discovered that the churning of their friends profoundly influences long-term producers. Last but not least, we demonstrated the predictive models based on the insights derived from this thesis could successfully predict whether a long-term producer will leave the site. The classifiers learned with the proposed feature sets verified the effectiveness of those features.

Future Works. There are many interesting directions that deserve further research. First of all, engagement patterns of newcomers using their location trajectories would be an important direction to study. We want to extract robust properties from user types, newcomers and long-term users, and develop advanced deep learning models to detect potential long-term users among newcomers. This research can help site maintainers to manage their user base from the influx of newcomers effectively. Second, analyzing user engagement using both users’ reviews and check-in information can enhance our understanding of user behaviors in LBSNs. A thorough analysis of users’ reviews and check-ins together would provide us with a more holistic view of users’ geographical
engagement patterns, which would lead us to exciting insights about human mobility. Finally, we want to incorporate more information such as demographic and personality traits of users so that we can identify a primary factor for each user type to churn. For that, we want to perform a comprehensive user survey with different types of users to investigate various motivations to stop contributing to the service.
REFERENCES


[83] Canwen Xu, Jing Li, Xiangyang Luo, Jiaxin Pei, Chenliang Li, and Donghong Ji. DLocRL: A Deep Learning Pipeline for Fine-Grained Location Recognition and Link-


This thesis has resulted in the publications of/in peer-reviewed venues:


Below is the list of other publications/submissions that are related to ubiquitous and mobile computing and user behavior analytics on online social media but not written into any chapter of this thesis:


Young D. Kwon, Jagmohan Chauhan, Abhishek Kumar, Pan Hui, and Cecilia Mascolo.
A Comprehensive Study of Lifelong Learning on Sensing-based Tasks. Submitted to the
ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT/UbiComp).
(Under Revision)