



University Transportation Research Center - Region 2

Final Report



Inferring High-Resolution Individual's Activity and Trip Purposes with the Fusion of Social Media, Land Use and Connected Vehicle Trajectories

Performing Organization: State University of New York (SUNY)



November 2017



Sponsor:
University Transportation Research Center - Region 2

University Transportation Research Center - Region 2

The Region 2 University Transportation Research Center (UTRC) is one of ten original University Transportation Centers established in 1987 by the U.S. Congress. These Centers were established with the recognition that transportation plays a key role in the nation's economy and the quality of life of its citizens. University faculty members provide a critical link in resolving our national and regional transportation problems while training the professionals who address our transportation systems and their customers on a daily basis.

The UTRC was established in order to support research, education and the transfer of technology in the field of transportation. The theme of the Center is "Planning and Managing Regional Transportation Systems in a Changing World." Presently, under the direction of Dr. Camille Kamga, the UTRC represents USDOT Region II, including New York, New Jersey, Puerto Rico and the U.S. Virgin Islands. Functioning as a consortium of twelve major Universities throughout the region, UTRC is located at the CUNY Institute for Transportation Systems at The City College of New York, the lead institution of the consortium. The Center, through its consortium, an Agency-Industry Council and its Director and Staff, supports research, education, and technology transfer under its theme. UTRC's three main goals are:

Research

The research program objectives are (1) to develop a theme based transportation research program that is responsive to the needs of regional transportation organizations and stakeholders, and (2) to conduct that program in cooperation with the partners. The program includes both studies that are identified with research partners of projects targeted to the theme, and targeted, short-term projects. The program develops competitive proposals, which are evaluated to insure the most responsive UTRC team conducts the work. The research program is responsive to the UTRC theme: "Planning and Managing Regional Transportation Systems in a Changing World." The complex transportation system of transit and infrastructure, and the rapidly changing environment impacts the nation's largest city and metropolitan area. The New York/New Jersey Metropolitan has over 19 million people, 600,000 businesses and 9 million workers. The Region's intermodal and multimodal systems must serve all customers and stakeholders within the region and globally. Under the current grant, the new research projects and the ongoing research projects concentrate the program efforts on the categories of Transportation Systems Performance and Information Infrastructure to provide needed services to the New Jersey Department of Transportation, New York City Department of Transportation, New York Metropolitan Transportation Council, New York State Department of Transportation, and the New York State Energy and Research Development Authority and others, all while enhancing the center's theme.

Education and Workforce Development

The modern professional must combine the technical skills of engineering and planning with knowledge of economics, environmental science, management, finance, and law as well as negotiation skills, psychology and sociology. And, she/he must be computer literate, wired to the web, and knowledgeable about advances in information technology. UTRC's education and training efforts provide a multidisciplinary program of course work and experiential learning to train students and provide advanced training or retraining of practitioners to plan and manage regional transportation systems. UTRC must meet the need to educate the undergraduate and graduate student with a foundation of transportation fundamentals that allows for solving complex problems in a world much more dynamic than even a decade ago. Simultaneously, the demand for continuing education is growing – either because of professional license requirements or because the workplace demands it – and provides the opportunity to combine State of Practice education with tailored ways of delivering content.

Technology Transfer

UTRC's Technology Transfer Program goes beyond what might be considered "traditional" technology transfer activities. Its main objectives are (1) to increase the awareness and level of information concerning transportation issues facing Region 2; (2) to improve the knowledge base and approach to problem solving of the region's transportation workforce, from those operating the systems to those at the most senior level of managing the system; and by doing so, to improve the overall professional capability of the transportation workforce; (3) to stimulate discussion and debate concerning the integration of new technologies into our culture, our work and our transportation systems; (4) to provide the more traditional but extremely important job of disseminating research and project reports, studies, analysis and use of tools to the education, research and practicing community both nationally and internationally; and (5) to provide unbiased information and testimony to decision-makers concerning regional transportation issues consistent with the UTRC theme.

Project No(s):

UTRC/RF Grant No: 49198-46-28

Project Date: November 2017

Project Title: Inferring High-Resolution Individual's Activity and Trip Purposes with the Fusion of Social Media, Land Use and Connected Vehicle Trajectories

Project's Website:

<http://www.utrc2.org/research/projects/inferring-high-resolution-individual-activity>

Principal Investigator(s):

Qing He

Assistant Professor

Department of Civil, Structural and Environmental Engineering

University at Buffalo

Buffalo, NY 14260

Tel: (716) 645-3470

Email: qinghe@buffalo.edu

Performing Organization(s):

State University of New York (SUNY)

Sponsor(s):

University Transportation Research Center (UTRC)

To request a hard copy of our final reports, please send us an email at utrc@utrc2.org

Mailing Address:

University Transportation Research Center

The City College of New York

Marshak Hall, Suite 910

160 Convent Avenue

New York, NY 10031

Tel: 212-650-8051

Fax: 212-650-8374

Web: www.utrc2.org

Board of Directors

The UTRC Board of Directors consists of one or two members from each Consortium school (each school receives two votes regardless of the number of representatives on the board). The Center Director is an ex-officio member of the Board and The Center management team serves as staff to the Board.

City University of New York

Dr. Robert E. Paaswell - Director Emeritus of NY
Dr. Hongmian Gong - Geography/Hunter College

Clarkson University

Dr. Kerop D. Janoyan - Civil Engineering

Columbia University

Dr. Raimondo Betti - Civil Engineering
Dr. Elliott Sclar - Urban and Regional Planning

Cornell University

Dr. Huaizhu (Oliver) Gao - Civil Engineering
Dr. Richard Geddes - Cornell Program in Infrastructure Policy

Hofstra University

Dr. Jean-Paul Rodrigue - Global Studies and Geography

Manhattan College

Dr. Anirban De - Civil & Environmental Engineering
Dr. Matthew Volovski - Civil & Environmental Engineering

New Jersey Institute of Technology

Dr. Steven I-Jy Chien - Civil Engineering
Dr. Joyoung Lee - Civil & Environmental Engineering

New York Institute of Technology

Dr. Nada Marie Anid - Engineering & Computing Sciences
Dr. Marta Panero - Engineering & Computing Sciences

New York University

Dr. Mitchell L. Moss - Urban Policy and Planning
Dr. Rae Zimmerman - Planning and Public Administration

(NYU Tandon School of Engineering)

Dr. John C. Falocchio - Civil Engineering
Dr. Kaan Ozbay - Civil Engineering
Dr. Elena Prassas - Civil Engineering

Rensselaer Polytechnic Institute

Dr. José Holguín-Veras - Civil Engineering
Dr. William "Al" Wallace - Systems Engineering

Rochester Institute of Technology

Dr. James Winebrake - Science, Technology and Society/Public Policy
Dr. J. Scott Hawker - Software Engineering

Rowan University

Dr. Yusuf Mehta - Civil Engineering
Dr. Beena Sukumaran - Civil Engineering

State University of New York

Michael M. Fancher - Nanoscience
Dr. Catherine T. Lawson - City & Regional Planning
Dr. Adel W. Sadek - Transportation Systems Engineering
Dr. Shmuel Yahalom - Economics

Stevens Institute of Technology

Dr. Sophia Hassiotis - Civil Engineering
Dr. Thomas H. Wakeman III - Civil Engineering

Syracuse University

Dr. Baris Salman - Civil Engineering
Dr. O. Sam Salem - Construction Engineering and Management

The College of New Jersey

Dr. Thomas M. Brennan Jr - Civil Engineering

University of Puerto Rico - Mayagüez

Dr. Ismael Pagán-Trinidad - Civil Engineering
Dr. Didier M. Valdés-Díaz - Civil Engineering

UTRC Consortium Universities

The following universities/colleges are members of the UTRC consortium under MAP-21 ACT.

City University of New York (CUNY)
Clarkson University (Clarkson)
Columbia University (Columbia)
Cornell University (Cornell)
Hofstra University (Hofstra)
Manhattan College (MC)
New Jersey Institute of Technology (NJIT)
New York Institute of Technology (NYIT)
New York University (NYU)
Rensselaer Polytechnic Institute (RPI)
Rochester Institute of Technology (RIT)
Rowan University (Rowan)
State University of New York (SUNY)
Stevens Institute of Technology (Stevens)
Syracuse University (SU)
The College of New Jersey (TCNJ)
University of Puerto Rico - Mayagüez (UPRM)

UTRC Key Staff

Dr. Camille Kamga: *Director, Associate Professor of Civil Engineering*

Dr. Robert E. Paaswell: *Director Emeritus of UTRC and Distinguished Professor of Civil Engineering, The City College of New York*

Dr. Ellen Thorson: *Senior Research Fellow*

Penny Eickemeyer: *Associate Director for Research, UTRC*

Dr. Alison Conway: *Associate Director for Education/Associate Professor of Civil Engineering*

Nadia Aslam: *Assistant Director for Technology Transfer*

Nathalie Martinez: *Research Associate/Budget Analyst*

Andriy Blagay: *Graphic Intern*

Tierra Fisher: *Office Manager*

Dr. Sandeep Mudigonda, *Research Associate*

Dr. Rodrigue Tchamna, *Research Associate*

Dr. Dan Wan, *Research Assistant*

Bahman Moghimi: *Research Assistant;*
Ph.D. Student, Transportation Program

Sabiheh Fagigh: *Research Assistant;*
Ph.D. Student, Transportation Program

Patricio Vicuna: *Research Assistant*
Ph.D. Candidate, Transportation Program

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Inferring High-Resolution Individual's Activity and Trip Purposes with the Fusion of Social Media, Land Use and Connected Vehicle Trajectories		5. Report Date 9/1/16-11/30/17	6. Performing Organization Code
7. Author(s) Qing He		8. Performing Organization Report No.	
9. Performing Organization Name and Address SUNY at Buffalo 17 Capen Hall, Buffalo, NY 14260-1660		10. Work Unit No.	11. Contract or Grant No. 49198-46-28
12. Sponsoring Agency Name and Address UTRC The City College of New York 137 th Street and Convent Avenue New York, NY 10031		13. Type of Report and Period Covered Final	
15. Supplementary Notes		14. Sponsoring Agency Code	
<p>16. Abstract</p> <p>Trip purpose is crucial to travel behavior modeling and travel demand estimation for transportation planning and investment decisions. However, the spatial-temporal complexity of human activities makes the prediction of trip purpose a challenging problem. With the increasing advance of the Information Communication Technology (ICT), tremendous social media data becomes available. The goal of this report is to model and predict trip purpose with social media data.</p> <p>In order to achieve the goal of this report, first, this report provides a new approach to match Point of Interests (POIs) from Google Places API with Twitter data. Therefore, the popularity of each POI can be obtained. Moreover, a Bayesian Neural Network is employed to model the trip dependence within each individual's daily trip chain and infer the trip purpose. In addition, to tackle the computational challenge in BNN, Elastic Net is implemented for feature selection before classification task. In addition, we also propose a Dynamic Bayesian Network for modeling and predicting trip purpose.</p> <p>Major findings are summarized as follows: We introduce a novel information retrieval method to match tweet with nearby Google Place Points of Interests (POIs) for trip prediction. The results show that our proposed method can reach up to 90% accuracy, whereas Foursquare tweet based method can only acquire 2%~16% accuracy.</p> <p>This study purposes a Dynamic Bayesian Network to model and predict trip purpose. Extensive experiments were conducted on real-world data sets, this method can achieve approximate 64% in average accuracy. This algorithm is more accurate when predict "shopping" activities, and the accuracy can be achieved as high as 80%.</p> <p>This report implements a feature selection method with Elastic Net. Total 29 features out of 45 are selected for modeling. The feature selection procedure is essential in a sense that it remarkably reduces the running time of BNN by 75%, from 60 minutes to 15 minutes.</p>			
17. Key Words Dynamic Bayesian Network, Trip Purpose, Bayesian Neural Network		18. Distribution Statement	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No of Pages 46	22. Price

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. The contents do not necessarily reflect the official views or policies of the UTRC or the Federal Highway Administration. This report does not constitute a standard, specification or regulation. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Table of Contents

List of Tables	1
List of Figures	2
EXECUTIVE SUMMARY	3
1 Introduction	6
2 Literature Review	9
2.1 Trip Purpose Inference and Prediction	9
2.2 Social Media Analytics and Applications in Transportation	9
2.3 Bayesian Neural Network	11
3 Data Description and preprocessing	12
3.1 Survey Data	12
3.2 Twitter Data	15
4 Methodology	17
4.1 Overview	17
4.2 POI Popularity Modeling	20
4.3 Dynamic Bayesian Network	24
4.3.1 Dynamic Bayesian Network Construction	24
4.3.2 Dynamic Bayesian Network Parameter Learning	25
4.3.3 Dynamic Bayesian Network Prediction	27
4.4 Bayesian Neural Network	27
4.5 Elastic Net	29
5 Experiments	33
5.1 POI Mention Extraction from Geo-tagged Tweets	33
5.2 DBN results	34
5.3 Feature selection results	36
5.4 BNN results	38
6 Conclusion	42
7 References	43

LIST OF TABLES

Table 4-1 Activity Categories	18
Table 4-2 POI Category	23
Table 5-1 Accuracy Comparison for Proposed POI Extraction Method and Foursquare-formatted Tweets	34
Table 5-2 Performance of DBN	35
Table 5-4 Performance of BNN	39

LIST OF FIGURES

Figure 3-1 Distribution of the number of activities. Note that we have deleted the days with only one “Home” activity. So, all the individuals in database at least have three activities after filtering. Including two “home” and one other activity.	14
Figure 3-2 Heat map of trip ends	15
Figure 3-3 Heat map of Geo-tagged tweets	16
Figure 4-1 A typical user's daily trips	18
Figure 4-2 The Dynamic Bayesian Network	24
Figure 4-3 Pseudocode of DBN Parameter Learning	27
Figure 4-4 Structure of a 2-layer Neural Network.....	28
Figure 5-2 Average Performance of Trip Purpose Inference.....	36

EXECUTIVE SUMMARY

Trip purpose is crucial to travel behavior modeling and travel demand estimation for transportation planning and investment decisions. However, the spatial-temporal complexity of human activities makes the prediction of trip purpose a challenging problem. With the increasing advance of the Information Communication Technology (ICT), tremendous social media data becomes available. The goal of this report is to model and predict trip purpose with social media data.

In order to achieve the goal of this report, first, this report provides a new approach to match Point of Interests (POIs) from Google Places API with Twitter data. Therefore, the popularity of each POI can be obtained. Moreover, a Bayesian Neural Network is employed to model the trip dependence within each individual's daily trip chain and infer the trip purpose. In addition, to tackle the computational challenge in BNN, Elastic Net is implemented for feature selection before classification task. In addition, we also propose a Dynamic Bayesian Network for modeling and predicting trip purpose.

Major findings are summarized as follows:

1. We introduce a novel information retrieval method to match tweet with nearby Google Place Points of Interests (POIs) for trip prediction. The results show that our proposed method can reach up to 90% accuracy, whereas Foursquare tweet based method can only acquire 2%~16% accuracy.
2. This study purposes a Dynamic Bayesian Network to model and predict trip purpose. Extensive experiments were conducted on real-world data sets, this method can achieve approximate 64% in average accuracy. This algorithm is more accurate when predict “shopping” activities, and the accuracy can be achieved as high as 80%.
3. This report implements a feature selection method with Elastic Net. Total 29 features out of 45 are selected for modeling. The feature selection procedure is essential in a sense that it remarkably reduces the running time of BNN by 75%, from 60 minutes to 15 minutes.

4. This study employs a Bayesian Neural Network to model trip purpose. And the BNN models outperform other prevailing algorithms. the BNN models can score the highest average accuracy which is almost 70%. Moreover, they can achieve extremely high performance for education activities which is approximate 92%, whereas other algorithms only reach low accuracy. As a result, BNN model is very powerful in the prediction of trip purpose.
5. It is found that trip duration is one of the most important features in the model. Therefore, the trip duration is a significant factor to infer the trip purpose. The accuracy will be deducted up to 9% without this feature.
6. The features from Google Places are also very significant. Further, the Twitter-related features will also improve prediction accuracy by itself. If we remove Google Places and Twitter-related features, the accuracy will decrease by 8% and 2%, respectively. However, if both are missing, the accuracy will decrease by almost 16%. Therefore, Google Places and Twitter data are essential to improve the accuracy of predicting trip purpose.

Major application areas are summarized as follows:

1. This can be utilized in **activity-based travel demand modeling**. Our method provides better results while predicting the trip purpose given a location. It can further enhance the accuracy of demand forecasting.
2. This can also be implemented in **survey labeling assistance**. Whenever a user finishes an activity, the survey labeling assistance service can give out three predictions, ordered by their probability. Then users can just choose the correct one instead of filling the survey.
3. Our research also can be applied to the **online recommendation**. Once the user inputs a destination in the online recommendation system, the proposed method can provide a prediction what the user might do in that location. Based on the predicted activity, we can recommend shops or display corresponding advertisements to the user.

To disseminate the outcome of this project, we have delivered two conference presentations: 2017 Joint Statistical Meeting (invited talk) and 2017 IEEE Big Data, one peer reviewed conference paper to be published in 2017 IEEE Big Data (acceptance rate 20%), and one journal paper submission to IEEE Transactions on Intelligent Transportation Systems (impact factor 3.72).

1 INTRODUCTION

Trip purpose is crucial to understanding travel behavior and estimating travel demand for transportation planning and investment decisions. Travel behaviors and travel patterns have become more and more temporally and spatially diverse and varied in past decades. This phenomenon is a synthetic consequence of a range of factors, for instance, increasing household car ownership, the growing number of double-income families, and the rising diversity of working time and places with the appearance of part-time work and work from home [1]. In order to capture this variation, a collection of reliable data is essential to acquire detailed travel information. Household travel survey is a widely used and traditional method. Conventional household travel surveys are composed of 1-day or multiple-day travel diaries of sampled households. Households are selected according to the proportion of certain groups of the same socio-demographic population. This subsample can reflect similar travel behaviors as the entire population. Even though household travel surveys are just a subsample of a population, this method is both high-expenditure and time consuming due to the large population base.

Trip purpose prediction was studied through several methods, and previous studies mainly used land use, temporal information, and socio-demographics obtained from household travel surveys. However, major challenges still widely exist. The existing methods simply lack detailed nearby Point of Interests (POIs) and the historical choices from other travelers.

Recently, ubiquitous various sensors can capture an enormous amount of passive temporal-spatial data. Each sensor has its own advantages and disadvantages. Compared to Call detail records (CDR) [2] [3], Wi-Fi, RFID [4], GPS is a better way to collect travel survey data since GPS devices can generate data every second with relatively high location accuracy. In this way, more reliable data can be collected and less burden imposed on participants.

On the other hand, social media is an emerging tool which allows people to create, share or exchange information and ideas by using texts, images or videos in a virtual community platform. Using social media to keep in touch with friends is a major method of communication among modern people. Since cell phones have developed to give people easy access to social media whenever they want, this enables users to generate spatial-temporal information immediately. Social media become more and more prevalent and ubiquitous. Not only young people are willing to share their moments and moods with their friends through social media, elder people are fond

of it to record their wonderful life and reconnect with friends for years [5]. Merchants (e.g. Yelp) offer varying degree of discount or gift when individuals check-in or post messages related to them makes users willing to provide location related information on social media. Individuals are also enthusiastic about checking in some smartphone apps (Foursquare Swarm) which can collect diverse medals or rewards and compete with their friends.

In addition to social media, Point of Interests (POIs) are favorable to extract land use information. They can be obtained from online location-based search and discovery services. For instance, Google Places Application Programming Interface (API) [6], searches POIs according to input place names or locations and search radius. The API will return detailed information about that place (category, location, opening hours and so on), or all detailed information of POIs within the search area of that location. With this method, we can not only query POIs in real time and identify trip purpose automatically, but also reduce the burden of individuals filling out traditional household travel surveys. In this report, we employ the function “nearby search” in Google Place API to extract POIs. “Nearby search” will return place names and place categories with respect to input coordinates and search radius. There are several characteristics about nearby search. There is no radius limitation for search requests, however, Google Places can only return at most 60 nearby POIs for the given geo-coordinates of each query, and a free Google Place API account has a limit of 150,000 free requests per 24 hours period.

During this research, we face three major challenges. First, users’ GPS records alone are not sufficient to infer their trip purposes. POIs near the trip ends can reveal land use, which can be related to the trip purpose. However, provided by a dozen of POIs, the POI information alone cannot capture users’ preferences on visiting the area. Therefore, we propose to mine social media data to help capture users’ popularity for each POI. The second challenge is how to extract POIs mentioned in Tweets. It is difficult to extract mentioned POIs from Tweets effectively because named entity extraction from short text is hard and social media data is very noisy. To address these two challenges, this report aims to predict (prior-trip)/infer (post-trip) the purpose of trip purposes (e.g., education, work, shopping, recreation, etc.) of individuals given their trip trajectories (or end locations), nearby POIs, and social media data. The third challenge is feature selection. The original dataset contains too many features and one cannot just throw all features into the model. Otherwise, the model will result in longer training time, lower accuracy, and more

irrelevant and noisy features.

The rest of this report is structured as follows: chapter 2 summarizes previous studies for trip purpose inference and the usage of social media. Chapter 3 introduces the datasets and the initial analysis result of the datasets. Chapter 4 presents the methodology. Chapter 5 discusses the results of the models. Finally, this report concludes in Chapter 6 by summarizing the main conclusions, potential applications, and future research.

2 LITERATURE REVIEW

2.1 Trip Purpose Inference and Prediction

We can extract abundant information directly from GPS data, such as trip start and end time, trip start and end location. This information can easily derive the origin-destination matrix which is very important for traffic demand modeling. However, travel mode and trip purpose cannot be obtained from original GPS data directly. Comparing to mining trip purpose from GPS data, travel mode is more straightforward [7]. Trip purpose is under-studied. Previous methods include rule-based methods [8, 9], probabilistic methods [10, 11], and machine learning and neural network methods [6, 12, 13]. Previous studies mainly used land use, temporal information, socio-demographics obtained from household travel survey. More literature can be found in Ermagun et al. [6]. Trip purpose can be separated into different categories. Three categories which are home, work/school, and other is the easiest and most common separate strategy. Alexander et al. divided trip purpose into home, work and other. Home is the maximum number of visit place from 7 pm to 8 am [14]. Work is the maximum distance (number of visit multiply distance for one place) from home, and other is otherwise. Increasing the number of categories provides more specific trip purpose. However, this also increases difficulties of inference and prediction. Despite different datasets utilized in different research, with Decision Tree algorithm, Deng achieved 87.6% accuracy when differentiating between 7 trip purposes [15]. Lu et al. obtained 73.4% accuracy when classifying 10 different trip purposes [16]. While for 12 distinguished trip purposes, overall accuracy that Oliveira et al. reached is only 65% [11]. Nevertheless, these accuracies are dominated by home or work trips on account of consist approximate one-third of total trips [6]. With the help of Google Places API which is a kind of online and real-time location-based query service, they achieved 67.14% overall prediction accuracy by using Random Forest algorithm when differentiating between 5 trip purposes other than home and work.

2.2 Social Media Analytics and Applications in Transportation

Since widely utilization of social media and passion of individuals generates a huge amount of passive data, social media data attracts attention in both academic and industrial area [17]. Yates et al. explored how social media technologies can influence emergency management on aspects of

knowledge sharing, reuse, and decision-making in an effective and efficient way [18]. Not only individuals share thoughts through social media, companies[19], merchants, agencies [20, 21], and even politicians[22] also utilize social media to carry their points. They broadcast information and influence individuals' opinions. Moreover, social media information can also be retrieved and benefit lots of fields including transportation.

Previous social media studies in transportation area mainly fall into two applications: traffic incident detection [23, 24], and traffic prediction[25-28]. Potential of social media in transportation have been exploring for several years. Several studies identified traffic accidents from tweets [21]. For example, Zhang et al. employed deep learning to detect traffic accidents from social media data [29]. The researchers considered that social media is an abundant, cost-effective, real-time data source which can complementary existing accident data source[30-32]. Others extracted traffic conditions (e.g. congestions, crash) from social media data which are helpful in traffic management and improve the level of service of traffic. Ni et al. utilized social media data to predict traffic flow and subway passenger flow under event conditions[33, 34]. Lin et al. modeled the impacts of inclement weather on freeway traffic speed by using social media data[35]. Zhang et al. explored the correlation between twitter concertation and traffic surge[36]. Social Media can also help understanding individuals travel behaviors. Zhang et al. researched the potentials of using social media to infer the longitudinal travel behavior by using a sequential model-based clustering method[37].

There are major challenges to be addressed before the use of tweets in extracting useful trip information. First, the tweet data is mainly comprised of the inherently complex and unstructured word texts, and the language ambiguity [38] in the tweet contents make them difficult to interpret. What's more, as the context of a tweet is limited to 140 words and tweet users usually intend to be concise, the common methods in the language processing studies such as support vector machine [31, 39], natural language processing [23, 40] maybe not be adaptive. Second, the information of activity location is embedded in the names that are not straightforward to be interpreted. For example, one cannot identify "paint and pour" as the "private school" until one searches it online. Third, same as other passively collected data, social media data generally lack ground truth of individuals' travel modes and trip purpose information. Inferring the ground truth of a personal trip is very difficult due to personal privacy. To address above challenges, this report employs

Google Place API to query trip end's categories and complement social media information.

2.3 Bayesian Neural Network

Deep learning is widely used and is state-of-the-art developments in all manner of data fields, and it can deal with massive dataset [41-44]. However, there are still some drawbacks to standard deep learning. First, neural networks compute point estimates for their weights, so that they make point predictions as well. Therefore, they may make overly confident about some classes. Second, a deep neural network has huge amount number of parameters. Therefore, it needs very large datasets in order to avoid overfitting. Third, in addition to the parameters the deep neural networks need to infer, there also have lots of configuration that need to be set, such as dropout probability, learning rate, network structure, and so on. Finally, Deep neural networks are poor at representing uncertainty.

Bayesian method is a way of updating probabilities or our beliefs about hypotheses given data. A probability distribution is the best way to represent the uncertainty of our knowledge and hypotheses. A Bayesian neural network is a neural network with a prior distribution on the weights. The history of Bayesian neural network can origin back to 30 years before. Denker et al. published a paper and hinted that integrating Bayesian over network parameters [45]. Then Bayesian neural networks entered its golden era around 1992. A series of papers that wrote by MacKay which described a quantitative and practical Bayesian framework for backpropagation networks [46]. Very few studies of Bayesian neural network can be found in the transportation field, Xie et al. employed Bayesian neural networks in transportation safety studies [47].

3 DATA DESCRIPTION AND PREPROCESSING

3.1 Survey Data

This report collects California Household Travel Survey (CHTS) data from February 2012 to January 2013. The geographical range of this survey covers all 58 counties of California and three adjacent counties in Nevada. In this report, we only utilize data within Bay Area. Two types of data, GPS data and survey data, are contained in CHTS dataset. In Bay Area, there are 108,778 individuals belongs to 42,431 households in this survey and completed one-day survey. For GPS data, 10,474 travelers from 5460 households carried GPS devices and reported 7 days of GPS data. Three types of GPS devices are implemented in this survey, wearable GPS device, in-vehicle GPS device, and in-vehicle GPS device plus an on-board diagnostic (OBD) unit. Each participant should fill a trip diary on a website or mail it to the institute. The GPS data records trip-related information, such as coordinates of the trip start and end locations, trip start and end times, trip modes, trip durations, and trip distances. The survey data includes activity related information, such as activity place names, activity purposes, and activity start and end times. In order to merge these two datasets together, several rules are created.

1. Survey activity data for each individual should contain at least two trips. Since first activity for all participants is to stay at home, they should have at least one activity other than home in order to generate a trip. Moreover, the last activity for all participants should also be 'home'.
2. At least one origin or destination of the trip in CHTS data should be located within Bay Area bounding box in order to cooperate with tweet data.
3. POI queried from Google Places API should be at least assigned to one POI category (money, leisure, food, bar, care, store, trans, auto, religion, civic, health, improve, edu, and lodge) since we need this POI to categorize trip ends.

There also several challenges in the data preprocessing procedure. First, participants fill in the inaccuracy or false travel information (e.g., round up arrive and departure times from activity locations, missing trip, wrong trip purpose, etc.). Therefore, information in survey data may not match information of GPS data perfectly. To address this issue, we first order one traveler's survey

data according to arrive time, activity and visit place sequence, and order this traveler's GPS data according to the trip end time, trip sequence and trip duration. Then we merge two datasets together. Second, one GPS trip may be divided into several records in GPS data when vehicles pass the tunnel or under other circumstances when the GPS signal is blocked. Moreover, several GPS trips are connected to one trip if the interval between each trip is too small, for instance, when one picks up or drop off passengers. After overcoming these challenges, we obtain total 43,767 activities which include 13,198 home activities and 30,569 non-home activities. The detail rule for categorizing trip purpose is shown in Table 4-2. Figure 3-1 depicts the distribution of activities conducted by individuals. Moreover, the average number of activities is 7.65 per day. According to first data filter rule mentions before, regardless the first and the last should happen at home, there still at least includes one activity other than home for all participants. Therefore, all the individuals in the database at least have three activities after filtering. Moreover, Figure 3-2 is the heat map of trip end locations.

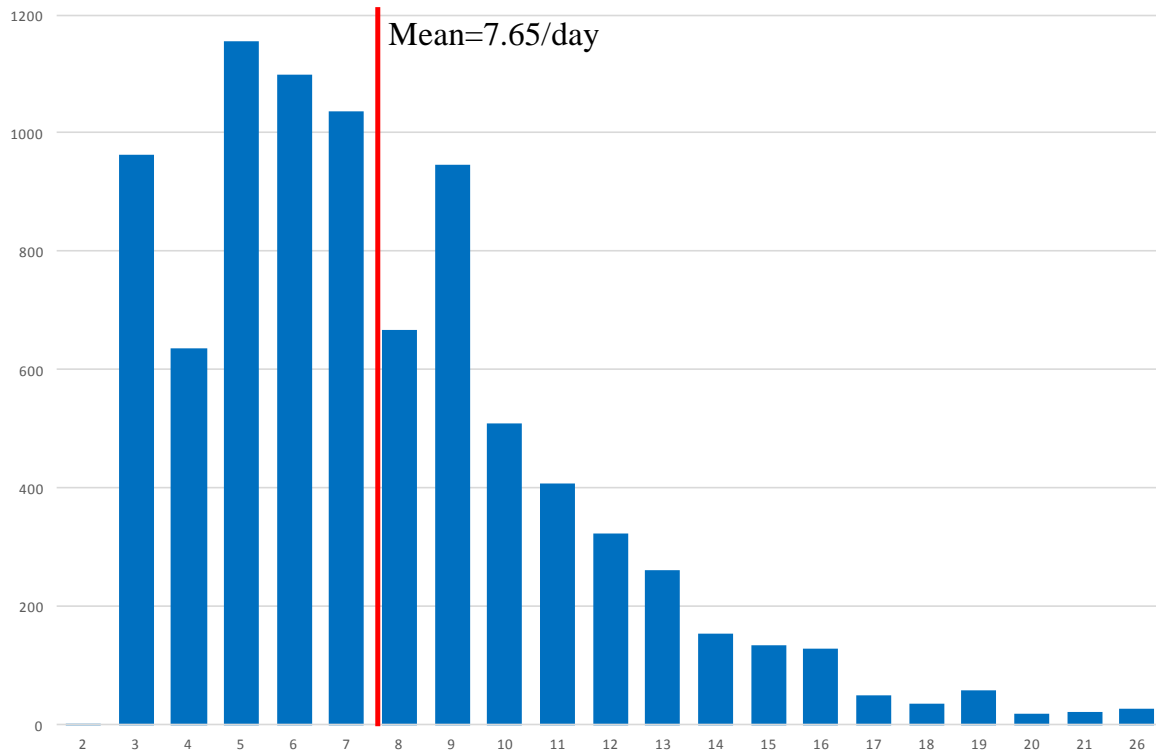


Figure 3-1 Distribution of the number of activities. Note that we have deleted the days with only one “Home” activity. So, all the individuals in database at least have three activities after filtering. Including two “home” and one other activity.

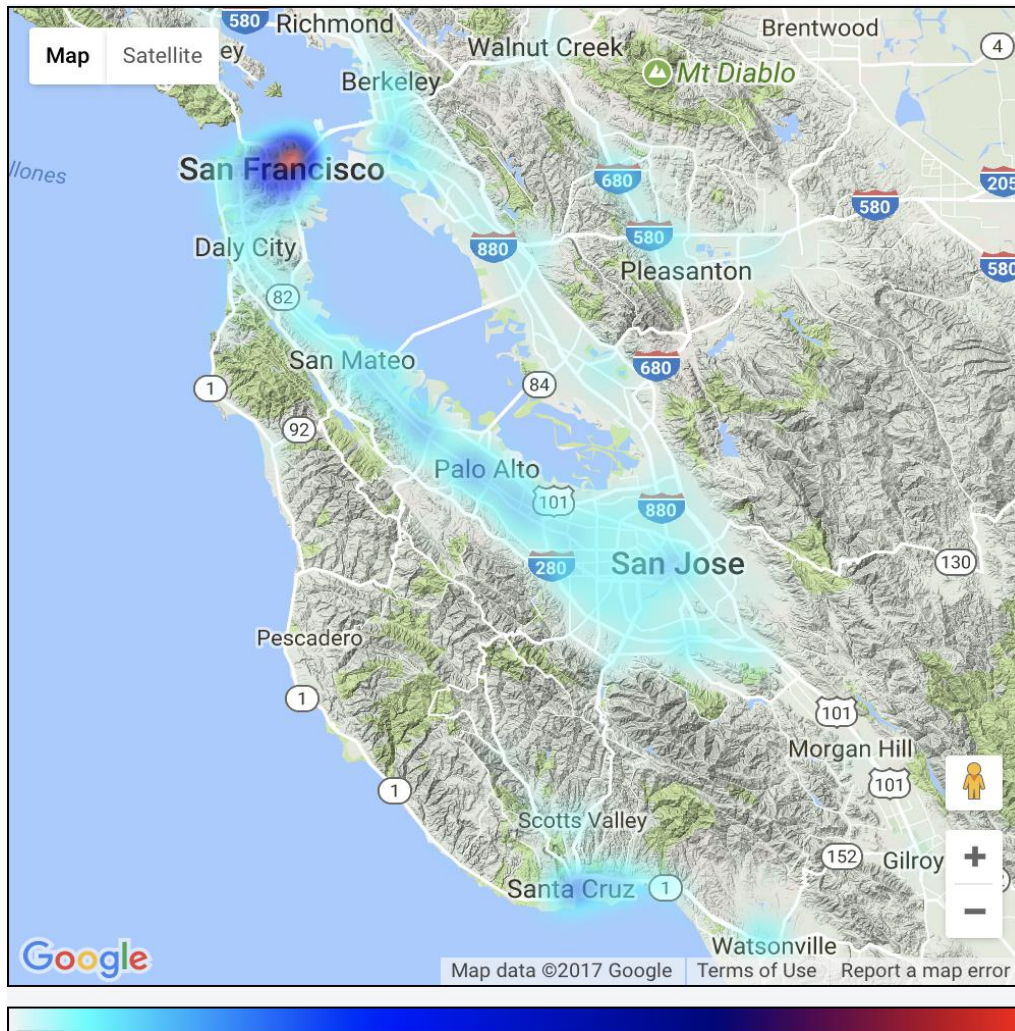


Figure 3-2 Heat map of trip ends

3.2 Twitter Data

Twitter data is collected from January 31st 2013 to February 16st 2017 in Bay Area with a bounding box (longitude: 121.75W~122.75W, latitude: 36.8N~37.8N). There are two types of tweet data which are geo-tagged tweets and non-geo-tagged tweets. In this report, we only use geo-tagged tweet data and there are near 6.9 million valid geo-tagged tweets in total. The heat map of geo-tagged tweets in Bay Area is shown in Figure 3-3

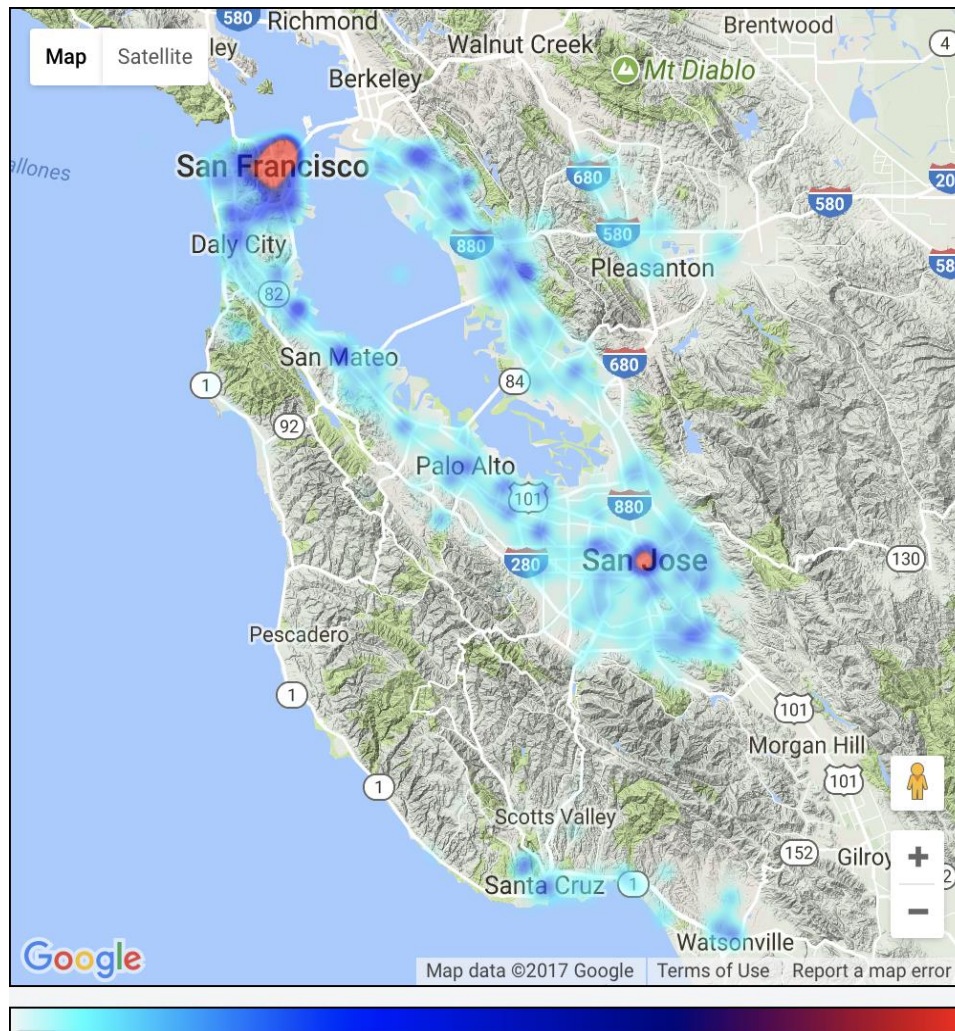


Figure 3-3 Heat map of Geo-tagged tweets

Since social media data is very noisy, it is very difficult to extract the named entity from a short text. In this report, we develop an information retrieval procedure to recognize mentioned POIs from nearby geo-tagged tweets. We will discuss it in Chapter 4.2.

4 METHODOLOGY

4.1 Overview

In the following, we introduce several important concepts that will be used throughout the work, then formally define the trip purpose inference problem.

Definition 1. A Trajectory (Tr) is a sequence of time-ordered spatial points, $Tr: l_1 \rightarrow l_2 \rightarrow \dots \rightarrow l_n$ where each point l is represented by a pair of GPS coordinates, i.e., longitude and latitude.

In this report, we regard “trip” as the movement from one location to another, e.g., $l_1 \rightarrow l_2$, and we refer these GPS points l as “trip end locations”. Take the trajectory in Figure 4-1 as an example. The trajectory comprises eight GPS points which follows the sequence of $l_{Home} \rightarrow l_{School} \rightarrow l_{Work} \rightarrow l_{Grocery} \rightarrow l_{Grocery} \rightarrow l_{Home} \rightarrow l_{Restaurant} \rightarrow l_{Home}$. From the perspective of trips, it has seven trips labelled by blue circles. However, in most cases, we cannot know the activities performed in a location without users’ input. In other words, we don’t know whether a user is shopping in a grocery store or having meal at a restaurant, given only the geo-coordinates of trip end locations.

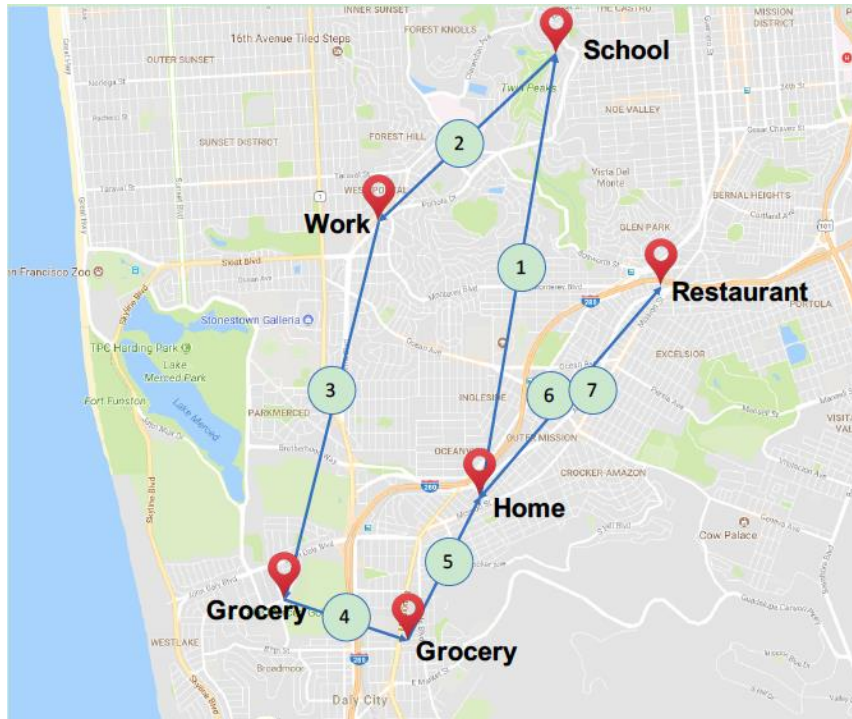


Figure 4-1 A typical user's daily trips

Definition 2. Trip Purpose is the activity that a user performed at a trip end location.

As the name refers, it denotes the purpose of a trip. In the following sections, we use “trip purpose” and “activity” interchangeably. As shown in Table 4-1, we categorize all trip purposes into eight categories, i.e., “Home”, “Education”, “Shopping”, “EatOut”, “Recreation”, “Personal”, “Work”, and “Transportation”.

Table 4-1 Activity Categories

Category	Example Activities
Home	Any activities performed at home
Education	School, Glass, Laboratory, Meal at college, After-school sports activities, Library, Clubs, etc.
Shopping	Routine shopping (groceries, clothing, convenience store, etc.), Shopping for major purchases or specialty items (appliance, electronics, new vehicles, major household repairs, etc.), Service private vehicle (gas, oil, lubes, etc.)
EatOut	Drive through meals (snacks, coffee, etc.), Eat meal at restaurant/dinner
Recreation	Indoor or outdoor exercise, Sports, Health care
Personal	Household errands, Personal business (visit attorney, accountant, etc.), Civic/religious activities, Entertainment (movies, sporting events, etc.)

Work	All activities performed at the work place
Transportation	Change type of transportation (walk to bus, walk to parking lot, etc.), Pick up/drop off passengers

Definition 3. Point of Interest (POI) is a specific location that someone may find useful. In this work, they represent venues in the physical world, e.g., banks and shopping malls. Each POI is associated with properties such as name, address, coordinates, category and etc.

Definition 4. Geo-tagged Tweet is a Twitter message associated with a pair of GPS coordinates where the message was generated.

Problem Definition. Given the trajectories of users, the points of interest and the Twitter messages near trip end locations, our Objective is to infer the purposes of trips.

Note that some of the trips have labels with corresponding purposes. The labels can be provided by users through Apps, manually recorded by users' trip diary, or mined from Twitter messages. However, acquiring these labels is very difficult, and it is usually assumed to be unavailable for a large portion of trips.

A trip's purpose is determined by many factors, such as other activities of the day, the category of the visited venue, the functionality and the popularity of the destination area. Before discussing the proposed methods, we first shed some light on how these factors associate with the trip purpose inference.

Sequential activities of the day. Common sense tells us that users' activities usually follow some patterns, and there are intrinsic relationships among the sequence of activities. For instance, parents may drop off their children at school before going to work, people would eat in a restaurant after shopping in a mall, and etc. Similar patterns among sequential activities widely exist, and it is very useful information for the purpose inference.

The category of the visited venue usually correlates with the trip purpose. For example, people arriving at a restaurant are very likely to have lunch or dinner; checking in at a mall tells us he will be shopping. There will be close relations between the category of venue people visited and their trip purposes. Unfortunately, the GPS devices are not accurate enough to pinpoint the venues

visited. In addition, it is also not easy to acquire this knowledge from people because of privacy concerns.

The functionality of the trip end area reveals the general usage of the nearby area. When we are not aware of the specific venue that a user visited, the nearby POIs can give us a hint about what the trip purpose would be. For example, arriving at a place with many shops nearby means people will go shopping with a higher probability. Specifically, the distribution of POI categories is a good feature to denote the functionalities of a location.

The popularity of the trip end area. Although the nearby POIs can help us understand the functionalities of a location, it cannot capture how people think about this area. Obviously, not all the venues attract equal attentions. In other words, some of them are more popular than the others. The popularity of the venues can be a useful feature for the purpose inference task, because the trip purpose is indeed a people-centric concept. Fortunately, social media can help us out here. Take Twitter as an example, people can send geo-tagged messages, and many of them contains the comments towards nearby POIs. By matching these geo-tagged tweets to real-world POIs, we can reveal venues' popularities accordingly.

A good trip purpose inference method needs to consider all these factors and the intrinsic relationships among them. In the following sections, we will first describe the proposed method for POI popularity modeling with social media data, then demonstrate the proposed Dynamic Bayesian Network approach.

4.2 POI Popularity Modeling

Based on the above reasoning, a POI's popularity can be captured if we can accurately identify them from tweet messages. The basic idea is that a POI is more popular if it has been mentioned by more tweets. However, this is a very challenging task. Firstly, social media data are very short. Existing named entity extraction methods perform poorly on these messages which have limited contexts. For example, "apple" may refer to the IT company or the fruit. Second, social media data are also very noisy. People usually use informal languages and names in tweets, such that we cannot expect to match a POI's full name in tweets. For instance, a tweet

“dinner 2 tacos from lacorneta” mentions a restaurant “La Corneta Taqueria” without the full name, but rather with an abbreviation.

In this section, we propose a method to learn POIs’ popularities from geo-tagged tweets. It can be much easier to identify mentioned POIs from these tweets because their associated geo-coordinates give us a good hint. Specifically, we can narrow down the search space to all nearby POIs. Compared with traditional methods [48] that works with a large POI knowledge base, our method can restrict the POI candidates to several dozens in a local area.

The proposed method works as follows. For each geotagged tweet, we first construct a local candidate pool with nearby POIs. Then a match index is calculated between the tweet content and each candidate POI names, and the POI with the highest index is marked as matched. Finally, the POI popularity of a trip end area can be derived by aggregating all the identified POIs from geo-tagged tweets. By this means, we can identify the mentioned POIs from geo-tagged tweets both effectively and efficiently. In the following, we describe each component in detail.

POI Local Candidate Pool Construction. In order to identify POI mentions for a geo-tagged tweet, we first construct a local candidate pool with all the POIs near the geo-coordinates of the tweet. Considering the accuracy of GPS devices, we shouldn’t set the range to be too small or too large. In this work, we set the range as 200 meters which usually results in a pool with several dozen candidates.

Calculate POI Match Index. After constructed a local candidate pool for each tweet, the next step is to find the best matched POI among all candidates. To this end, we design a match index to measure the similarity between a tweet and a POI name. This match index considers two essential factors:

- The number of matched terms. The match index should be larger if there are more terms matched between a tweet and a POI name. For example, a tweet “Was just told by a teenager working at this Jamba Juice, that I looked like a young Walter White” mentions the POI “Jamba Juice Redwood City”, and two terms are matched between them. However, there is another nearby POI “Geoff White Photographers” which

matches a term “white” to the tweet. In this case, “Jamba Juice” with 2 matched terms should be weighed higher than the other one with only 1 matched term.

- The rareness of the matched terms. Some terms may frequently appear in the candidate pool. For example, there is no surprise that many POI names contain “San Francisco” in the Bay area. Then these terms should have less impact on the matching index. On the contrary, terms such as “Corneta” is relatively rare. In fact, this term only appears in a restaurant named “La Corneta Taqueria”. No doubt that these terms should have larger impact on the matching index. In other words, if terms like “Corneta” matched between a tweet and a POI name, we should have a high belief that the tweet mentioned the restaurant “La Corneta Taqueria”.

In this work, we propose a POI Match Index which characterizes the aforementioned factors. Specifically, each Tweet T_i is represented by a set of terms, i.e., $T_i = \{u_1, u_2, \dots, u_m\}$. Similarly, each candidate POI’s name is represented by $P_j = \{v_1, v_2, \dots, v_n\}$. Then the set of Matched Terms MT between T_i and P_j are

$$MT(T_i, P_j) = T_i \cap P_j \quad (4-1)$$

We can calculate the Match Index as follows

$$MI(T_i, P_j) = |MT(T_i, P_j)| \times \log \frac{N_{pool}}{1 + \sum_{k=1}^N \mathbb{1}(MT(T_i, P_j) \in P_k)} \quad (4-2)$$

where N_{pool} denotes the size of the candidate pool. $\mathbb{1}(\cdot)$ is the indicator function which returns 1 if and only if the condition holds. In addition, $\sum_{k=1}^N \mathbb{1}(MT(T_i, P_j) \in P_k)$ calculates the frequency of the matched terms MT in the POI candidate pool. As shown in the equation 2, the first term considers the number of matched terms, and the second term considers the rareness of the matched terms. Note that $|MT(T_i, P_j)|$ could be zero, i.e., there are no matched terms between T_i and P_j . In this case, the Match Term Index will be 0 which is reasonable.

After calculated the Match Index between T_i and every P_j in its candidate pool, we can return the one with the highest index as the identified POI. However, we still need to set a threshold to the MI, and return nonidentified if none of MIs exceed the threshold. In sum, With the constructed local POI candidate pool and the proposed Match Term Index, we can accurately identify the nearby POIs that mentioned in the Tweets. The POI with the highest match index is returned.

POI Popularity Modeling. After extracted the POIs mentions from social media data, we can further represent the POI popularity across different categories by counting the corresponding mentions from social media. For example, if restaurants are mentioned by 10 different tweets, we will count 10 towards the popularity of the POI category “Food”. Then the counts can be normalized into a distribution across all the POI categories in Table 4-2.

Table 4-2 POI Category

POI Category	Google Place Type
Auto	car repair, car wash, gas station
Bar	bar, night club
Care	beauty salon, hair care, spa
Civic	courthouse, lawyer, police, fire station, city hall, embassy, local government, local, city hall, embassy, local government, local, government office
Edu	library, school, university
Food	bakery, cafe, food, meal takeaway, restaurant, meal delivery
Health	dentist, doctor, health, hospital, pharmacy, physiotherapist, veterinary care
Improve	electrician, locksmith, moving company, painter, plumber, real estate agency, painter, plumber, real estate agency, travel agency, general contractor, roofing, contractor, insurance agency, laundry, storage
Leisure	amusement park, aquarium, art gallery, casino, bowling alley, gym, movie rental, movie theater, museum, park, stadium, zoo
Lodge	rv park, lodging, campground
Money	accounting, atm, bank, post office, finance

Religion	cemetery, church, funeral home, hindu, temple, mosque, place of worship, synagogue
Store	convenience store, department store, electronics store, florist, furniture store, grocery, supermarket, grocery or supermarket, hardware store, home goods, store, jewelry store, liquor store, pet store, jewelry store, liquor store, pet store
Trans	bus station, subway station, train station, taxi stand, parking, car rental, airport, light rail station, transit station

4.3 Dynamic Bayesian Network

4.3.1 Dynamic Bayesian Network Construction

In this work, we propose a Dynamic Bayesian Network (DBN) to model people’s sequential activities. As shown in the Figure 4-2, $a \in A$ denotes the activity performed (or trip purpose), $c \in C$ denotes the category of POI a user visited, and l is the trip end locations. All activities (A) and POI categories (C) defined in this report are shown in Table 4-1 and Table 4-2.

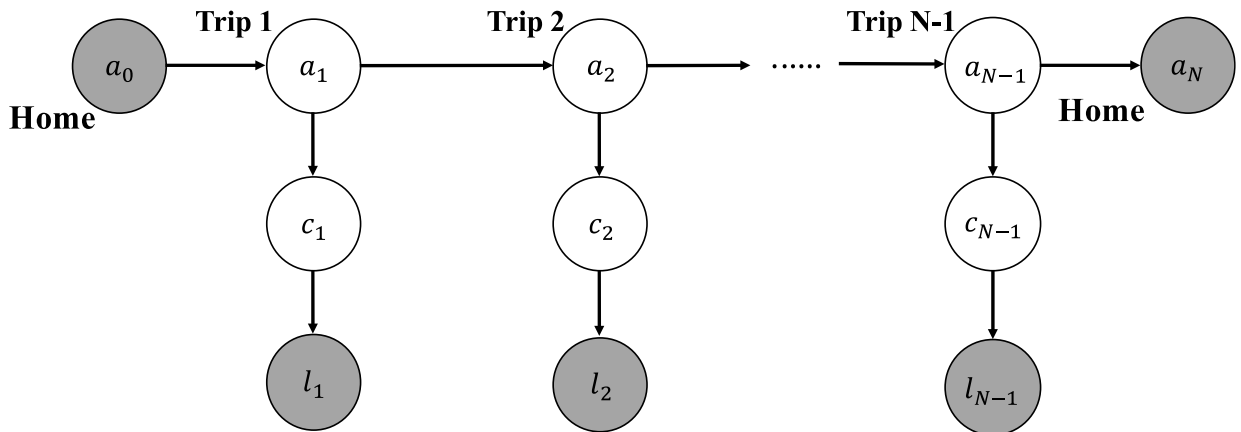


Figure 4-2 The Dynamic Bayesian Network

The DBN model can be interpreted in a generative process, or in other words, in a user’s decision making process. For each trip i , a user first decides an activity a_i (or purpose) based on his previous one a_{i-1} . Then he chooses a venue category c_i based on this choice of activity. At last, he chooses a geo-location l_i to finally perform the activity a_i in venue c_i . This process continues until the last trip.

The likelihood function of the proposed DBN model is as follows

$$P(a, c, l) = P(a_0)P(c_0|a_0)P(l_0|c_0) \times (\prod_{i=1}^N P(a_i|a_{i-1})P(c_i|a_i)P(l_i|c_i)) \quad (4-12)$$

where $P(a_i|a_{i-1})$ is the probability of the activity a_i given previous activity a_{i-1} , $P(c_i|a_i)$ is the probability of the visited POI category given current activity a_i , and $P(l_i|c_i)$ is the probability of chosen location l_i given currently chosen POI category c_i . In the proposed model, the visited location l_i is always observed, such that we can use Bayes's rule to approximate $P(l_i|c_i)$ as follows:

$$P(l_i|c_i) \propto \frac{P(c_i|l_i)}{P(c_i)} \propto \frac{P_{POI}(c_i|l_i)P_{tweet}(c_i|l_i)}{\sum_i P_{POI}(c_i|l_i)P_{tweet}(c_i|l_i)} \times \frac{1}{P(c_i)} \quad (4-13)$$

In the Equation 4-13, $(c_i|l_i)$ denotes the POI category distribution given a geo-location l_i . This distribution is determined by two aforementioned factors: the functionality distribution $P_{POI}(c_i|l_i)$, and the popularity distribution $P_{tweet}(c_i|l_i)$. The first distribution is obtained from populating the nearby POIs, and the second distribution is obtained by extracting POI mentions from nearby tweets.

4.3.2 Dynamic Bayesian Network Parameter Learning

There are two sets of parameters in the DBN model: the transition probabilities $P(a_i|a_{i-1})$ and the emission probabilities $P(c_i|a_i)$. Note that in our problem, the activities a_i and visited venues c_i are not fully observed. In other words, many activities and corresponding venues are not labelled in the data. In order to learn the parameters from such incomplete data, we adopt the EM algorithm[54]. The process is summarized in Figure 4-3. It starts with an initial set of parameters. In each Expectation step (E-step), we compute the expected sufficient statistics for the parameter variables. Then in each Maximization step (M-step), we treat the expected sufficient statistics as observed, and perform Maximum Likelihood Estimation to estimate a new set of parameters. The algorithm continues between these two steps until converges.

Algorithm 2 DBN Parameter Learning

```

1  Input: DBN Structure  $\mathcal{G}$ , Partially observed trip dataset  $\mathcal{D}$ , Initial set of Parameters  $\theta_0 =$ 
    $\{P(a_i|a_{i-1}), P(c_i|a_i)\}$ 
2  Output: Learned DBN parameters  $\theta_t$ 
3  for  $t \leftarrow 0, 1, \dots$ , until convergence do
4      // Expectation-step
5      for each  $a \in A$  and each  $c \in C$  do       $\triangleright$ Initialize
6           $M_t[\square_i, a_{i-1}] = 0$ 
7           $M_t[c_i, a_i] = 0$ 
8      end for
9      for each  $d \in D$  do
10         Run inference on the graph  $\mathcal{G}$  using evidence  $d$ 
11         for each  $a \in A$  and each  $c \in C$  do
12              $M_t[a_i, a_{i-1}] \leftarrow M_t[a_i, a_{i-1}] + P(a_i, a_{i-1}|d)$ 
13              $M_t[c_i, a_i] \leftarrow M_t[c_i, a_i] + P(c_i, a_i|d)$ 
14         end for
15     end for
16     // Maximization-step
17     for each  $a \in \square$  and each  $c \in C$  do
18          $P(a_i|a_{i-1}) \leftarrow \frac{M_t[a_i, a_{i-1}]}{M_t[a_{i-1}]}$ 
19          $P(c_i|a_i) \leftarrow \frac{M_t[c_i, a_i]}{M_t[a_i]}$ 
20          $\theta_{t+1} \leftarrow \{P(a_i|a_{i-1}), P(c_i|a_i)\}$ 
21     end for
22 end for
23 return  $\theta_t$ 

```

Figure 4-3 Pseudocode of DBN Parameter Learning

4.3.3 *Dynamic Bayesian Network Prediction*

With the learned parameters of the DBN mode, we can infer possible activities and their corresponding probabilities for any given trip. Specifically, we can calculate the posterior probability of the j th activities given a user's trajectory $d \in D$. This estimates the probability of activity a_j out of all possible activities A , as shown in Equation 4-14.

$$P(a_j|d) = \frac{P(a_j,d)}{P(d)}, \forall a_j \in A \quad (4-14)$$

The returned results are possible activities ranked by their probabilities. Note that, generating a ranked result is a great advantage by adopting the Bayesian method, especially compared with traditional methods which can only provide a best guess. In fact, the top ranked inference results are very useful in real-world applications. Many classification tasks may have very vague decision boundaries, and usually the best guess results in poor performance. However, a ranked list with probabilities can help us identify several meaningful results and improve the inference accuracy. The experiments shown in Chapter 5-4 provides a good demonstration.

4.4 **Bayesian Neural Network**

Before discussing the Bayesian neural network, we introduce the neural network first. A neural network can be represented as a weighted directed graph in which are nodes and directed edges with weights are connections between neuron outputs and neuron inputs. The weighted inputs are all summed up inside neurons and passed through an activation function. The activation function is a set of transfer functions used to scale the summations to the desired value. A neural network contains three kinds of layers, an input layer, an output layer, and hidden layers. In this report, we build a neural network with an input layer, an output layer, and two hidden layers, which is also called multi-layer perceptron, and the structure is shown in Figure 4-4. The activation function between input layer and hidden layer, and between two hidden layers are tanh functions [49]. And between second hidden layer and output layer, the activity function is a sigmoid function [50].

The essential goal is to obtain a distribution over the prediction. In order to achieve this goal, the posterior needs to be sampled lots of time randomly. Actually, BNN sums over lots of neural networks with slightly different weights, the weights are sampled from the posterior. And the Bayesian neural network is an ensemble form of a lot of neural networks. The predictive distribution of output y^* given a new input x^* is shown as follow,

$$p(y^*|x^*, X, y) = \int p(y^*|x^*, w)p(w|W, y)dw \quad (4-4)$$

BNN involves estimation of each distribution of the parameter. Therefore, the entire process is much more computationally expensive than the simple point estimation. To tackle this issue, feature selection will choose a subset of significant features and greatly reduce the model training time in this case. Also, a smaller number of features is more interpretable than overmuch features.

4.5 Elastic Net

We first introduce Least absolute shrinkage and selection operator (LASSO), a prevailing regularization and feature selection method in statistic and machine learning [51]. LASSO uses L_1 norm for regularization, which minimize the summation of squared errors subjected to an upper bound on the summation of absolute value of model parameters. The object function of LASSO is shown as follow:

$$\min_{\beta} ||Y - X\beta|| \quad s. t. \quad \sum_{j=1}^p |\beta_j| \leq t \quad (4-5)$$

Ridge regression is another method of regularization which employs L_2 norm [52]. Ridge regression is aiming to estimate $\hat{\beta}$ will minimizes the sum of squared error and satisfy the following constrain:

$$\min_{\beta} (Y - \beta^T X)^T (Y - \beta^T X) + \lambda \beta^T \beta \quad s. t. \quad \sum_{j=1}^p \beta_j^2 \leq c \quad (4-6)$$

Both LASSO and ridge regression have advantages and disadvantages. LASSO cannot deal with the data that the number of features is greater than the number of dataset entries. Moreover, for highly correlated features, LASSO tends to select one feature from each group of correlated features. LASSO encourages shrinking of coefficients to 0 and dropping those variables from your

model. On the other hand, ridge regression can handle the case that the number of features is greater than the number of records. Ridge regression tends to keep all variables. This is not desirable in the feature selection task.

Elastic Net is another alternative to complete the feature selection task [53]. Elastic Net is a hybrid method which mixes LASSO and ridge regression together, and it is trained with L_1 and L_2 norm as regularizer. It can take advantage of both LASSO and ridge regression and overcome their shortcomings. We start to introduce Naïve Elastic Net (NEN). The objective function is shown below:

$$\hat{\beta}(NEN) = \arg \min_{\beta} (Y - \beta^T X)^T (Y - \beta^T X) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \quad (4-7)$$

This equation can be rewritten as a penalized least square:

$$\hat{\beta} = \arg \min_{\beta} |Y - X\beta|^2 \quad \text{subject to } (1 - \alpha) \|\beta\|_1 + \alpha \|\beta\|_2^2 \leq t \quad (4-8)$$

with $\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2}$ and $(1 - \alpha) \|\beta\|_1 + \alpha \|\beta\|_2^2$ the Elastic Net penalty. And α ranges from 0 to 1. Moreover, if $\alpha = 0$, the procedure becomes LASSO. If $\alpha = 1$, the process becomes the ridge regression. The NEN is just a simple combination of LASSO and ridge regression, and this process does not perform satisfactorily. The reason is that there are two shrinkage procedure in this process. A ridge shrinkage followed by a lasso shrinkage does not decrease variance much but increase the bias of the estimate. Therefore, overall prediction errors increase. It is found that the rescaled NEN will provide better performance, and it is called Elastic Net [53]. The relationship between NEN and Elastic Net is shown below.

$$\hat{\beta}(Elastic\ Net) = (1 + \lambda_2) \times \hat{\beta}(NEN) \quad (4-9)$$

This process will undo shrinkage. Moreover, for orthogonal design matrix, the LASSO solution is minimax optimal. Elastic Net will achieve the same minimax optimality after rescaling by $(1 + \lambda_2)$.

In order to find the solution for Elastic Net, we should solve NEN problem first, then rescale it. For a fixed λ_2 , the NEN problem is equivalent to a LASSO problem. And LASSO already has

an efficient algorithm to solve it called Least Angle Regression (LARS). The Pseudocode for LARS is shown as follow in .

Algorithm 1 LARS

- 1 **Input:** $X_{N \times P}, Y_{N \times 1}$
 - 2 **Output:** coefficients $\theta_{P \times 1}$
 - 3 Set all coefficient $\theta_i (i = 1, \dots, P)$ equal to 0
 - 4 Set active set $A = \emptyset$
 - 5 Find predictor x_{jm} that correlated with $Y_{N \times 1}$ the most
 - 6 Let direction $D = x_{jm}$
 - 7 **Repeat:** Adjust the coefficient in the direction D at the highest step possible until some other explanatory variable x_{jm} has the same absolute correlation residual $r = Y - \hat{Y}$
 - 8 Put x_{jm} in A
 - 9 Let D in the direction that is equiangular with $x_{j1}, x_{j2}, \dots, x_{jm}$
 - 10 **Until:** $P - 1$ variables have entered the active set A
-

Figure 4-5 Pseudocode of LARS algorithm

After we find the solution of NEN, Elastic Net problem is also solved. In this report, our problem is a multiclass classification task. Multinomial model is employed with Elastic Net to location appropriate feature subset. The multinomial model and the Elastic Net penalized negative log-likelihood function are shown below,

$$\Pr(C = k | X = x) = \frac{e^{\beta_{0k} + \beta_k^T x}}{\sum_{l=1}^K e^{\beta_{0l} + \beta_l^T x}} \quad (4-10)$$

$$\begin{aligned} l(\{\beta_{0k}, \beta_k\}_1^K) = & - \left[\frac{1}{N} \sum_{i=1}^N \left(\sum_{k=1}^K y_{ik} \left(e^{\beta_{0k} + \beta_k^T x} \right) - \log \sum_{k=1}^K e^{\beta_{0k} + \beta_k^T x} \right) \right] + \lambda [(1 - \\ & \alpha) \sum_{j=1}^p ||\beta_j|| + \alpha ||\beta||_2^2 / 2] \end{aligned} \quad (4-11)$$

This is a K levels $C = \{1, 2, \dots, K\}$ multinomial problem. Let X be a $N \times p$ input matrix, and Y be a $N \times K$ indicator response matrix, with elements $y_{il} = I(c_i = l)$. β is a $p \times K$ matrix of coefficients. β_k refers to the k th column for outcome category k , and β_j is the j th row for vector of K coefficients for variable j . It allows only (β_{0k}, β_k) to vary for a single category at a time. For each value of λ , it first iterates over all categories indexed by k , computes each time a quadratic approximation about the parameters of the current class. Then the inner process is an inner loop which uses a quadratic approximation to the log-likelihood, and the coordinate descent on the resulting penalized weighted least-squares problem.

5 EXPERIMENTS

5.1 POI Mention Extraction from Geo-tagged Tweets

In Section III-A, we propose to extract mentioned POIs from geo-tagged tweets. For each geo-tagged tweet, we can accurately identify the mentioned POI with a local candidate pool and the match index. In practice, it is very hard to evaluate the performance of POI mention extraction from tweets, because we have so many nearby tweets and trips in the data set. Figure 5-1 shows the histogram of tweets near trip ends. There are 26,593 out of 30,569 (87%) non-home activities have nearby tweets. The median of nearby tweet number is 294 per activity location, and the mean is 2607 per activity location, respectively. Therefore, the distribution of number nearby activity location of tweets is a heavy-tailed distribution. Most non-home activity locations have a large number of geo-tagged tweets.

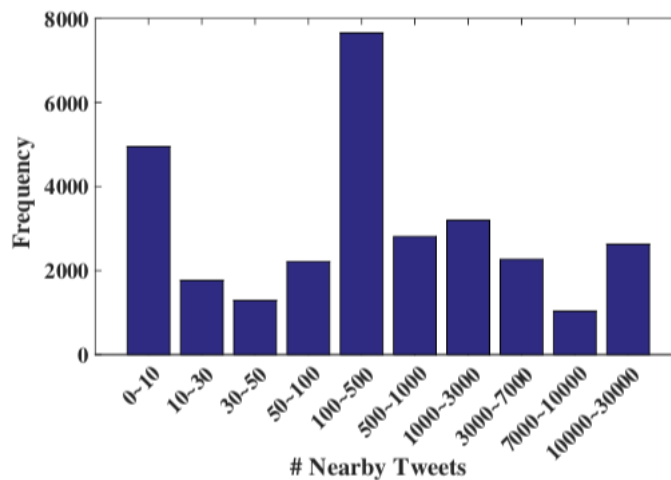


Figure 5-1 Histogram of tweets near trip ends

Actually, it is impossible to be evaluated without a standard data set labelled by human workers. To this end, we recruited volunteers to label the POI mentioned tweets near 50 random trip end locations. Given a list of tweets, the volunteers are asked to judge whether those tweets mentioned any nearby POI, and whether the identified POIs are correct. In Table 5-1, we present the results on several example trip end locations and the average performance. For each location, we populate the total number of nearby tweets, the number of POI-mentioned tweets which are identified by

human workers. Some of the tweets are formatted by third party Apps, such as Foursquare and Instagram. We can easily parse POIs from these well formatted tweets, for example, “Im at Applewood Pizza in San Carlos Ca” and “Bagel time! @ Bagel Street Cafe Town Center Alameda”. However, there are still many POI-mentioned tweets without these formats. For instance, “Catch us at amc Mercado 2012 am insurgent!!!!”. The better performance shown in Table 5-1 indicates the proposed method can also extract mentioned POIs from free tweets.

Table 5-1 Accuracy Comparison for Proposed POI Extraction Method and Foursquare-formatted Tweets

Location	Total nearby tweets	POI-mentioned tweets	Foursquare-formatted tweets	Proposed method extracted tweets
Location 1	394	131	3 (2.2%)	120 (91.6%)
Location 2	338	134	8 (6.0%)	108 (80.6%)
Location 3	1928	210	6 (2.9%)	152 (72.3%)
Location 4	562	245	16 (6.5%)	210 (85.7%)
Location 5	7172	3158	527 (16.7%)	2428 (76.9%)
Location 6	10927	4465	455 (10.1%)	3637 (81.5%)

5.2 DBN results

To compare the performance of all the methods on the trip purpose inference, we conduct experiments on the collected real-world data set. The features used for training include travel mode, previous activity category, activity time and duration, nearby POI category distribution, and nearby POI popularity distribution. We randomly select 80% trips as training data, and leave the rest for testing. All the methods are evaluated by 10 times, and the average results are reported. In addition, there are about one third trip end locations in the dataset are either home or work, because the survey collected trips from people’s daily lives. Using these trips to train the model would greatly bias the performance. As a result, we only test the inference results on other activities, and assume

the home and work trips are known. In other words, we are interested in non-trivial tasks of inferring non-home and non-work trip purposes.

Note that the proposed DBN model can output a ranked list of activities with probabilities. This is a great advantage by adopting the Bayesian method. We also evaluate the DBN performance with top-2 and top-3 inference results in the experiment, denoted as DBN-top-2, and DBN-top-3 respectively. In these cases, we regard the inference as correct if the ground truth activity is among the top-2 or top-3 results.

Table 5-2 Performance of DBN

Accuracy	SVM	ANN	KNN	RF	DBN-top-1	DBN-top-2	DBN-top-3
EatOut	4.80%	30.00%	32.10%	60.60%	79.00%	83.20%	85.50%
Personal	7.70%	21.70%	22.50%	50.40%	42.20%	65.00%	90.00%
Recreation	18.60%	39.00%	30.60%	55.50%	62.70%	77.40%	84.70%
Education	27.00%	41.90%	34.40%	40.20%	52.30%	72.90%	78.40%
Shopping	59.60%	65.30%	52.50%	78.60%	80.10%	94.20%	98.40%
Transportation	84.60%	74.20%	59.10%	75.40%	68.30%	84.30%	89.60%
Average	33.70%	45.40%	38.50%	60.10%	64.10%	79.50%	87.80%
F1	SVM	ANN	KNN	RF	DBN-top-1	DBN-top-2	DBN-top-3
EatOut	0.087	0.349	0.32	0.635	0.712	0.835	0.879
Personal	0.134	0.299	0.266	0.55	0.476	0.737	0.92
Recreation	0.281	0.43	0.342	0.585	0.592	0.721	0.826
Education	0.362	0.461	0.358	0.419	0.484	0.757	0.84
Shopping	0.517	0.6	0.452	0.756	0.754	0.853	0.909
Transportation	0.625	0.655	0.588	0.722	0.735	0.869	0.919
Average	0.334	0.465	0.387	0.611	0.626	0.795	0.882

The inference accuracy and F1 scores are shown in Table 5-2, and the average results are also compared in Figure 5-2. We can observe that the proposed DBN model, including DBN-top-1, DBN-top-2 and DBN-top-3, outperforms other baselines on almost every activity category by

higher accuracy and F1 score. This is because DBN model captures the intrinsic relationships among sequential activities, trip end locations' POI distributions and the popularities identified from the Twitter data. On average, the DBN model can reach 64% accuracy with the top-1 inference result. Moreover, the accuracy of top-2 and top-3 ranked results can reach 79.5% and 87.8%. These results are impressive and they demonstrate that the top-ranked results generated by the DBN model is very useful in the trip purpose inference.

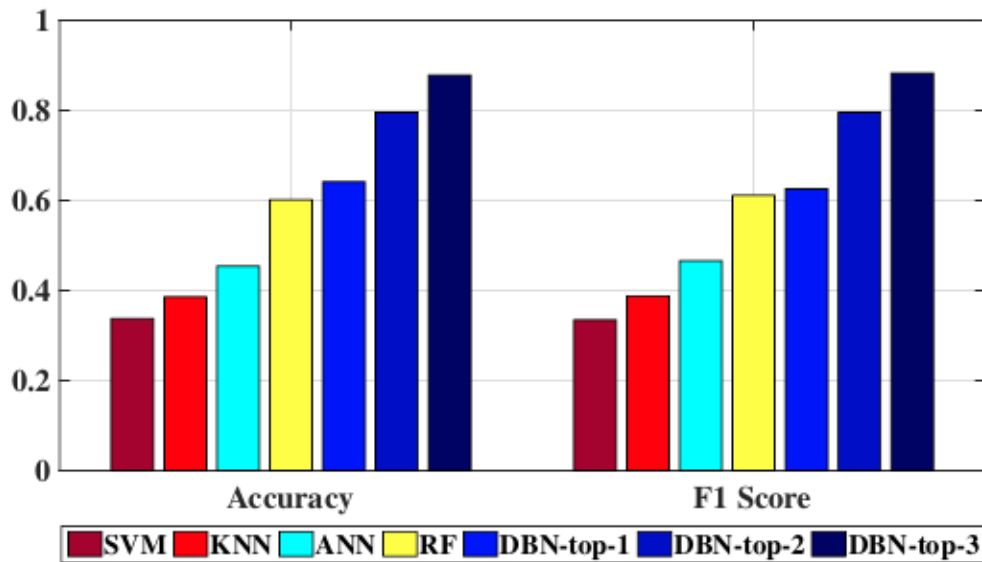


Figure 5-2 Average Performance of Trip Purpose Inference

5.3 Feature selection results

In this report, we have 45 features, and more than 20,000 records in total. In order to achieve great feature selection result, we first need to determine the value of α in Equation 4-11. α ranges from 0 to 1. And $\alpha = 0$ is LASSO, $\alpha = 1$ is ridge regression respectively. We iterate 21 different α from 0 to 1 with a 0.05 step size. The experiment applies feature selection by using the Elastic Net with the Artificial Neural Network (ANN). The best α is set as 0.75 after this experiment. As a result, there are 29 features selected from 45 features, as shown in Table 5-3. As one can see, the number of features chosen is reasonable. These 29 features are across four categories, social-demographic, trip information, Google Places API information, and Twitter information. Therefore, this feature subset capture traveler, trip, and place information completely. Moreover,

this feature subset can achieve the highest accuracy with the ANN. Then this feature subset is used for BNN model training, and the running time of training process decreases from 60 minutes to 15 minutes with a 75% saving.

Table 5-3 Feature selection result of Elastic Net

Features from Elastic Net	Feature Description
Social-Demographic	
employment	if the traveler is an employee
student	if the traveler is a student
Age	the age of traveler
Trip Information	
depTime	the departure time of the trip
TripDistance	the trip distance to the destination
TripDuration	the trip duration to destination
home	if this traveler conducted home activity
person	if this traveler conducted personal business activity
rec	if this traveler conducted recreation activity
shop	if this traveler conducted shopping activity
work	if this traveler conducted work activity
trans	if this traveler conducted transit activity
Google Places API Information	
G_MONEY	how many money Google Places POIs near trip end location
G_FOOD	how many food Google Places POIs near trip end location
G_BAR	how many bar Google Places POIs near trip end location
G_CARE	how many care Google Places POIs near trip end location

G_TRANS	how many transit/transportation Google Places POIs near trip end location
G_AUTO	how many auto Google Places POIs near trip end location
G_CIVIC	how many civic Google Places POIs near trip end location
G_IMPROVE	how many improvement Google Places POIs near trip end location
G_EDU	how many education Google Places POIs near trip end location
G_RELIGION	how many religion Google Places POIs near trip end location
<hr/> Twitter Information <hr/>	
T_FOOD	how many food related tweets posted near trip end location
T_CARE	how many care related tweets posted near trip end location
T_STORE	how many store related tweets posted near trip end location
T_TRANS	how many transit/transportation related tweets posted near trip end location
T_RELIGION	how many religion related tweets posted near trip end location
T_IMPROVE	how many improvement related tweets posted near trip end location
T_LODGE	how many lodge related tweets posted near trip end location

5.4 BNN results

In this section, we further compare the BNN model with several state-of-art algorithms, including Support Vector Machine (SVM) [55], ANN, K-nearest Neighbors (KNN) [56], and random forest (RF) [57]. SVM aims to map data into a high dimension and construct a set of hyperplanes to divide data with maximum margins. ANN is introduced in the previous section, and we apply the same structure as the BNN model in this experiment. KNN tends to classify objects to the class of the majority vote of its neighbors. RF is an ensemble learning method that constructing numbers of decision trees together in the training stage and classification result is

defined as the mode of classes. Finally, accuracy and F1 score are employed to measure the performances, and the equation of F1 score is shown below.

$$Accuracy = \frac{true\ positive}{true\ positive + false\ positive} \quad (5-1)$$

$$F_1 = 2 \times \frac{1}{\frac{1}{recall} + \frac{1}{precision}} = 2 \times \frac{precision \times recall}{precision + recall} \quad (5-2)$$

In the dataset, there are more than 40% of trip purpose are home and work, and this observation complies with one’s daily life pattern. However, the model trained with this dataset may lead to great bias. As a result, we only utilize data without home and work trip purpose [6]. Moreover, for model validation, we implement 5-fold cross-validation and average the results.

One of the advantages of BNN is that it can provide the probability of being each category in trip purpose. We not only compare the top 1 result (the category with highest prediction probability) of the BNN model named BNN-top-1, but also explore if top 2 and top 3 predictions hit the correct category. We denote them as BNN-top-2 and BNN-top-3, respectively.

From Table 5-4, one can observe that the BNN models outperform other algorithms on almost every trip purpose category in accuracy and F1 score. However, the accuracy and F1 score of “shopping” and “personal” are both lower compared to others. The reason is that it is difficult to distinguish and define “personal business” and “shopping” activities. BNN, which estimates every parameter in the model as a distribution, provides a better description of data and leads to better accuracy. Therefore, the BNN models can score the highest average accuracy. Moreover, they can achieve extremely high performance for education activities, whereas other algorithms only reach low accuracy. As a result, BNN model is very powerful in the prediction of trip purpose.

Table 5-4 Performance of BNN

Accuracy	SVM	ANN	KNN	RF	BNN-top-1	BNN-top-2	BNN-top-3
EatOut	4.80%	30.00%	32.10%	60.60%	64.26%	74.38%	85.70%

Personal	7.70%	21.70%	22.50%	50.40%	48.17%	60.94%	92.43%
Recreation	18.60%	39.00%	30.60%	55.50%	62.30%	80.17%	89.11%
Education	27.00%	41.90%	34.40%	40.20%	91.83%	93.70%	95.49%
Shopping	59.60%	65.30%	52.50%	78.60%	64.77%	76.26%	89.36%
Transportation	84.60%	74.20%	59.10%	75.40%	84.85%	88.00%	91.06%
Average	33.70%	45.40%	38.50%	60.10%	69.36%	78.91%	90.52%
F1 Score	SVM	ANN	KNN	RF	BNN-top-1	BNN-top-2	BNN-top-3
EatOut	0.087	0.349	0.32	0.635	0.636	0.742	0.864
Personal	0.134	0.299	0.266	0.55	0.575	0.704	0.927
Recreation	0.281	0.43	0.342	0.585	0.616	0.771	0.883
Education	0.362	0.461	0.358	0.419	0.905	0.951	0.971
Shopping	0.517	0.6	0.452	0.756	0.649	0.746	0.900
Transportation	0.625	0.655	0.588	0.722	0.752	0.812	0.887
Average	0.334	0.465	0.387	0.611	0.689	0.788	0.905

Further, we conduct experiments on specific features to find out that how much the accuracy will decrease if such features are removed from modeling. Results are shown in Figure 5-3.

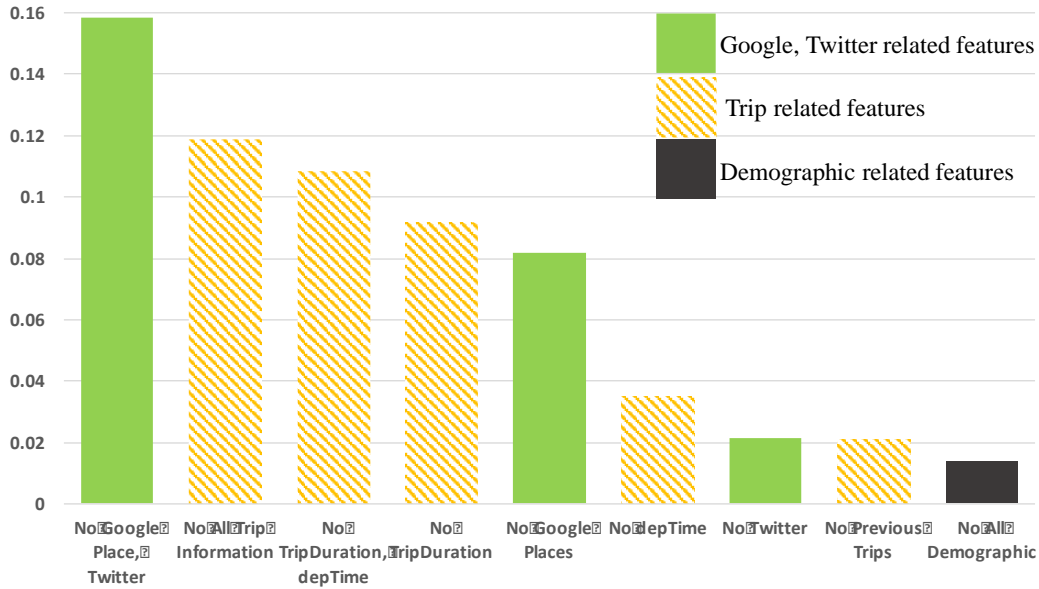


Figure 5-3 Accuracy deduction for removing different set of features

As one can see from Figure 5-3, if the model does not include features from Google Places and Twitter, the accuracy will drop the most. Then the followings are features from trip information and ones from the trip duration and departure time. It is found that trip duration is one of the most important features in the model. Therefore, the trip duration is a significant factor to infer the trip purpose. The features from Google Places are also very significant. Further, the Twitter-related features will also improve prediction accuracy by itself. If we remove Google Places and Twitter-related features, the accuracy will decrease by 8% and 2%, respectively. However, if both are missing, the accuracy will decrease by almost 16%. Therefore, Google Places and Twitter data are essential to improve the accuracy of predicting trip purpose.

6 CONCLUSION

First, we introduce a novel information retrieval method to match tweet with nearby Google Place Points of Interests (POIs) for trip prediction. The results show that our proposed method can reach up to 90% accuracy, whereas Foursquare tweet based method can only acquire 2%~16% accuracy.

Second, this report implements a feature selection method with Elastic Net. Total 29 features out of 45 are selected for modeling. The feature selection procedure is essential in a sense that it remarkably reduces the running time of BNN by 75%, from 60 minutes to 15 minutes.

Third, this study employs a Bayesian Neural Network to model trip purpose. And the BNN models outperform other prevailing algorithms. One major advantage of the Bayesian model is that it can return the possibility with each potential activity. The experiment shows a very high probability of correct prediction within the top 2 or top 3 ranked results.

Fourth, we find out that Google Places and Tweet greatly increase the accuracy compared with the model with other features. We also discover that trip duration can also greatly increase the accuracy of trip purpose inference.

Fifth, this study also purposes a Dynamic Bayesian Network to model and predict trip purpose. Extensive experiments were conducted on real-world data sets, and the results demonstrate advantages of the proposed method on accurately inferring the trip purposes.

Our research has following possible applications. First, this can be utilized in activity-based travel demand modeling. Our method provides better results while predicting the trip purpose given a location. It can further enhance the accuracy of demand forecasting. The second is survey labeling assistance. Whenever a user finishes an activity, the survey labeling assistance service can give out three predictions, ordered by their probability. Then users can just choose the correct one instead of filling the survey. Our research also can be applied to the online recommendation. Once the user inputs a destination in the online recommendation system, the proposed method can provide a prediction what the user might do in that location. Based on the predicted activity, we can recommend shops or display corresponding advertisements to the user. In the future, researchers can come up with better methods to mine more useful information from social media data, and this information can benefit the modeling and estimation of travel behavior.

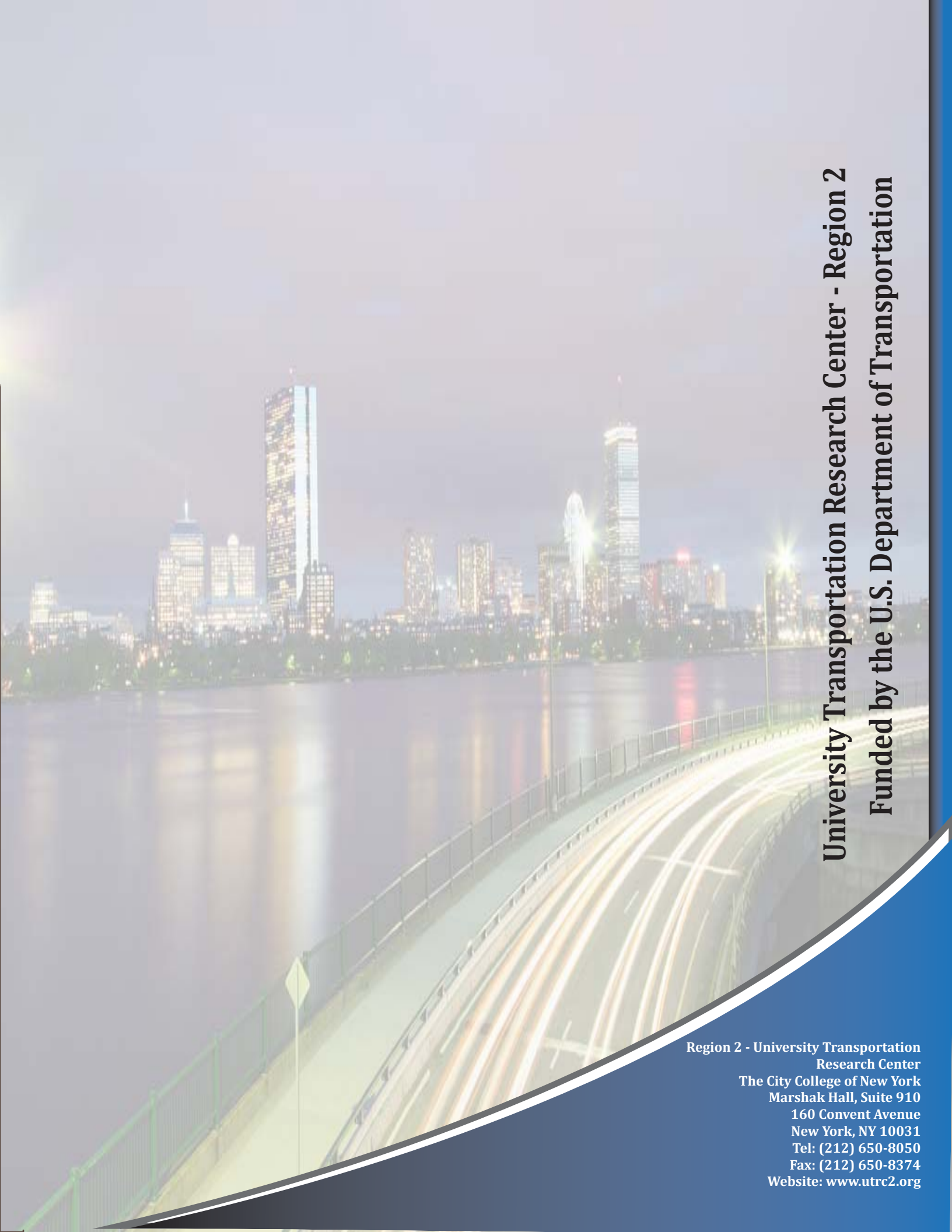
7 REFERENCES

- [1] W. Bohte and K. Maat, "Deriving and Validating Trip Destinations and Modes for Multiday GPS-Based Travel Surveys: Application in the Netherlands," in *Transportation research board 87th annual meeting*, 2008, no. 08-2268.
- [2] B. Cici, A. Markopoulou, E. Frias-Martinez, and N. Laoutaris, "Assessing the potential of ride-sharing using mobile and social data: a tale of four cities," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014, pp. 201-211: ACM.
- [3] N. B. Ponieman, A. Salles, and C. Sarraute, "Human mobility and predictability enriched by social phenomena information," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2013, pp. 1331-1336: ACM.
- [4] C. M. Schneider, C. Rudloff, D. Bauer, and M. C. González, "Daily travel behavior: Lessons from a week-long survey for the extraction of human mobility motifs related information," in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, 2013, p. 3: ACM.
- [5] M. Anderson and A. Perrin, "Tech adoption climbs among older adults. Pew Research Center," ed, 2017.
- [6] A. Ermagun, Y. Fan, J. Wolfson, G. Adomavicius, and K. Das, "Real-time trip purpose prediction using online location-based search and discovery services," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 96-112, 2017.
- [7] X. Su, H. Caceres, H. Tong, and Q. He, "Online travel mode identification using smartphones with battery saving considerations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 10, pp. 2921-2934, 2016.
- [8] J. Wolf, R. Guensler, and W. Bachman, "Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1768, pp. 125-134, 2001.
- [9] J. Wolf, S. Bricka, T. Ashby, and C. Gorugantua, "Advances in the application of GPS to household travel surveys," in *National Household Travel Survey Conference, Washington DC*, 2004.
- [10] C. Chen, H. Gong, C. Lawson, and E. Bialostozky, "Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study," *Transportation Research Part A: Policy and Practice*, vol. 44, no. 10, pp. 830-840, 2010.
- [11] M. Oliveira, P. Vovsha, J. Wolf, and M. Mitchell, "Evaluation of Two Methods for Identifying Trip Purpose in GPS-Based Household Travel Surveys," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2405, pp. 33-41, 2014.
- [12] L. Liao, D. Fox, and H. Kautz, "Extracting places and activities from gps traces using hierarchical conditional random fields," *The International Journal of Robotics Research*, vol. 26, no. 1, pp. 119-134, 2007.
- [13] G. Xiao, Z. Juan, and C. Zhang, "Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 447-463, 2016.

- [14] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin–destination trips by purpose and time of day inferred from mobile phone data," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 240-250, 2015.
- [15] Z. Deng and M. Ji, "Deriving rules for trip purpose identification from GPS travel survey data and land use data: A machine learning approach," in *Traffic and Transportation Studies 2010*, 2010, pp. 768-777.
- [16] Y. Lu and Y. Liu, "Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies," *Computers, Environment and Urban Systems*, vol. 36, no. 2, pp. 105-108, 2012.
- [17] Y. Lv, Y. Chen, X. Zhang, Y. Duan, and N. L. Li, "Social media based transportation research: the state of the work and the networking," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 1, pp. 19-26, 2017.
- [18] D. Yates and S. Paquette, "Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake," *International journal of information management*, vol. 31, no. 1, pp. 6-13, 2011.
- [19] M. J. Culnan, P. J. McHugh, and J. I. Zubillaga, "How large US companies can use Twitter and other social media to gain business value," *MIS Quarterly Executive*, vol. 9, no. 4, 2010.
- [20] F.-Y. Wang, "Scanning the issue and beyond: Real-time social transportation with online social signals," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 3, pp. 909-914, 2014.
- [21] X. Zheng *et al.*, "Big data for social transportation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 3, pp. 620-630, 2016.
- [22] M. Broersma and T. Graham, "Social media as beat: Tweets as a news source during the 2010 British and Dutch elections," *Journalism Practice*, vol. 6, no. 3, pp. 403-419, 2012.
- [23] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom, and P. Chaovalit, "Social-based traffic information extraction and classification," in *ITS Telecommunications (ITST), 2011 11th International Conference on*, 2011, pp. 107-112: IEEE.
- [24] E. Mai and R. Hranac, "Twitter interactions as a data source for transportation incidents," in *Proc. Transportation Research Board 92nd Ann. Meeting*, 2013, no. 13-1636.
- [25] A. Gal-Tzur, S. M. Grant-Muller, T. Kuflik, E. Minkov, S. Nocera, and I. Shoor, "The potential of social media in delivering transport policy goals," *Transport Policy*, vol. 32, pp. 115-123, 2014.
- [26] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," *Transportation research part C: emerging technologies*, vol. 68, pp. 285-299, 2016.
- [27] E. M. Daly, F. Lecue, and V. Bicer, "Westland row why so slow?: fusing social media and linked data sources for understanding real-time traffic conditions," in *Proceedings of the 2013 international conference on Intelligent user interfaces*, 2013, pp. 203-212: ACM.
- [28] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865-873, 2015.
- [29] Z. Zhang, Q. He, J. Gao, and M. Ni, "A Deep Learning Approach for Detecting Traffic Accidents from Social Media Data," *Transportation Research Part C*, vol. 86, pp. 580-596, 2018.

- [30] S. Grosenick, *Real-time traffic prediction improvement through semantic mining of social networks*. University of Washington, 2012.
- [31] A. Schulz, P. Ristoski, and H. Paulheim, "I see a car crash: Real-time detection of small scale incidents in microblogs," in *Extended Semantic Web Conference*, 2013, pp. 22-33: Springer.
- [32] Y. Gu, Z. S. Qian, and F. Chen, "From Twitter to detector: real-time traffic incident detection using social media data," *Transportation research part C: emerging technologies*, vol. 67, pp. 321-342, 2016.
- [33] M. Ni, Q. He, and J. Gao, "Using social media to predict traffic flow under special event conditions," in *The 93rd Annual Meeting of Transportation Research Board*, 2014.
- [34] M. Ni, Q. He, and J. Gao, "Forecasting the subway passenger flow under event occurrences with social media," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1623-1632, 2017.
- [35] L. Lin, M. Ni, Q. He, J. Gao, and A. W. Sadek, "Modeling the impacts of inclement weather on freeway traffic speed: exploratory study with social media data," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2482, pp. 82-89, 2015.
- [36] Z. Zhang, M. Ni, Q. He, J. Gao, J. Gou, and X. Li, "An Exploratory Study on the Correlation between Twitter Concentration and Traffic Surge 2," *Transportation Research Record*, vol. 35, p. 36, 2016.
- [37] Z. Zhang, Q. He, and S. Zhu, "Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 396-414, 2017.
- [38] P.-T. Chen, F. Chen, and Z. Qian, "Road traffic congestion monitoring in social media with hinge-loss Markov random fields," in *Data Mining (ICDM), 2014 IEEE International Conference on*, 2014, pp. 80-89: IEEE.
- [39] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, "Real-time detection of traffic from twitter stream analysis," *IEEE transactions on intelligent transportation systems*, vol. 16, no. 4, pp. 2269-2283, 2015.
- [40] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, "Tedas: A twitter-based event detection and analysis system," in *Data engineering (icde), 2012 ieee 28th international conference on*, 2012, pp. 1273-1276: IEEE.
- [41] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85-117, 2015.
- [42] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115-143, 2002.
- [43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
- [44] W. Dong, J. Li, R. Yao, C. Li, T. Yuan, and L. Wang, "Characterizing driving styles with deep learning," *arXiv preprint arXiv:1607.03611*, 2016.
- [45] J. Denker *et al.*, "Large automatic learning, rule extraction, and generalization," *Complex systems*, vol. 1, no. 5, pp. 877-922, 1987.
- [46] D. J. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural computation*, vol. 4, no. 3, pp. 448-472, 1992.

- [47] Y. Xie, D. Lord, and Y. Zhang, "Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis," *Accident Analysis & Prevention*, vol. 39, no. 5, pp. 922-933, 2007.
- [48] W. Hua, K. Zheng, and X. Zhou, "Microblog entity linking with social temporal context," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 2015, pp. 1761-1775: ACM.
- [49] Wikipedia. (2017, December 20). *Hyperbolic function*. Available: https://en.wikipedia.org/wiki/Hyperbolic_function
- [50] Wikipedia. (2017, 20 December 2017). *Sigmoid function*. Available: https://en.wikipedia.org/wiki/Sigmoid_function
- [51] Wikipedia. (2017, 20 December). *Lasso (statistics)*. Available: [https://en.wikipedia.org/wiki/Lasso_\(statistics\)](https://en.wikipedia.org/wiki/Lasso_(statistics))
- [52] Wikipedia. (2017, 20 December). *Tikhonov regularization*. Available: https://en.wikipedia.org/wiki/Tikhonov_regularization
- [53] Wikipedia. (2017, 20 December). *Elastic net regularization*. Available: https://en.wikipedia.org/wiki/Elastic_net_regularization
- [54] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. 2009.
- [55] Wikipedia. (2017, 2017). *Support vector machine*. Available: https://en.wikipedia.org/wiki/Support_vector_machine
- [56] Wikipedia. (2017, 20 December). *k-nearest neighbors algorithm*. Available: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
- [57] Wikipedia. (2017, 20 December). *Random forest*. Available: https://en.wikipedia.org/wiki/Random_forest

A long-exposure photograph of a city skyline at night, reflected in a body of water. In the foreground, a bridge or highway is visible with light trails from moving vehicles. The sky is dark, and the city lights are bright and colorful.

University Transportation Research Center - Region 2
Funded by the U.S. Department of Transportation

**Region 2 - University Transportation
Research Center**
The City College of New York
Marshak Hall, Suite 910
160 Convent Avenue
New York, NY 10031
Tel: (212) 650-8050
Fax: (212) 650-8374
Website: www.utrc2.org