The Study of Gender and User Mobility Features Using Twitter Data

Xinyi Liu [*1] **Mingren Shen** [*1] **Faust Shi** [*1]

Abstract

User travel patterns are widely studied using social media data like Twitter while not as systematic as using other data sources such as cell phone data and traditional survey data. Specifically, the relationships between people's demographic characteristics (e.g., gender, age and occupation) and their typical temporal and spatial travel patterns(e.g., periodicity of visiting a restaurant, time of leaving home on weekdays) are unclear. Therefore, this project tries to use temporal and spatial travel pattern features combined with Twitter text content features derived from more than 254,000 tweets to predict user's gender. The contribution of each feature to the model is weighed to measure its relative importance with the gender label. The test results show that several features are relatively more important, including users' tweet frequency on weekdays, tweet frequency on weekends, tweet frequency in the afternoon, travel distance between home and entertainment space and so on.

1. Introduction

Social media are popular platforms which can record a variety of users' personal information. Among this information, demographic characteristics (Sloan et al., 2013) and space-time travel patterns (Hasan et al., 2013) are two of great interesting areas to geographers. Demographic characteristics include gender, age, occupation and so on. Space-time travel patterns refer to the law of people's travel trajectories along both space and time (e.g., periodicity of visiting a restaurant, time of leaving home on weekdays). Traditionally, these two pieces of information are gathered from surveys and interviews (Chen et al., 2011). As the Information and Communication Technology (ICT) develops, they can be mined from cell phone records and other electronic GPS records (Siła-Nowicka et al., 2016). However, these methods either cost too much money and too large resources or encroach privacy. Nowadays, social media data could take part in these works with little expense and free of privacy infringement (Preoţiuc-Pietro & Cohn, 2013). Researchers have created robust framework and methodologies to mine space-time travel patterns from the geo-tagged online messages with temporal stamps. However, some direct demographic records(e.g., gender and occupation information on Twitter) is missing for some online platforms. It is important to design and implement methods to mine them from online messages (Rao et al., 2011).

It is important to restore user demographic information and link them with space-time travel patterns to unveil their relationships (Ahn et al., 2016). The results indicate social segmentation and contribute to the development of public facilities for different subgroups (Kang et al., 2010). However, few studies are conducted to investigate their relationships. Thus the travel patterns are discussed in terms of a general group of people, which is insufficient for human mobility studies because diverse travel patterns of different subgroups have been studied using other data sources (Kang et al., 2010). To fill the research gap, this paper tries to develop a machine learning classifier to study the characteristic space-time travel pattern features of different subgroups, specifically different genders (male/female/others). Test results of our classifier are compared with an online word-based Bayes network classifier and a first-name-based classifier. Our classifier shows improvement in either test accuracy or in test F1 scores.

2. Relate Work

2.1. Twitter Gender Inference

Automatically inferring user gender from Twitter is heavily investigated by both academic socitey and industry because gender is one of the most important demographic property of the user. Generally, there are 3 main approaches used for deriving Twitter users gender information: (1) profile-based (2) content-based (3) hybrid (Beretta et al., 2015).

2.1.1. Profile-based Gender Inference

Profile-based methods use the meta-data of the user's account in Twitter to help determine the gender of the users

[*]Equal contribution [1]University of Wisconsin,Madison , USA. Correspondence to: Xinyi Liu <xliu636@wisc.edu>.

CS 760 Final Project, UW-Madison, 2017.

(Sloan et al., 2015). In Twitter, a user can show his name, description, location, followers and friends publicly. Although Twitter does not check the authenticity of profile information, several studies have proven that most Twitter users provide their real name and real gender in their public profile (Cesare et al., 2017). The simplest and best feature of profile information is users' first name. Previous studies have shown that by comparing name record from nation demographic survey, first name based gender classifier can achieve real good performance (Sloan et al., 2013; Mislove et al., 2011) .There are several mature services like genderize.io and packages [1] [2] inferring gender using only first name. For example `genderize.io` provides API that can be used to determine gender of a first name with the help of a database contains 216286 distinct names across 79 countries and 89 languages. Generally, the profile-based method is considered as the benchmark of gender inference due to the high efficacy it can achieve. For example, Liu et al. use first name as the main feature to infer gender in Twitter and they obtain the accuracy around 85% (Liu & Ruths, 2013).

2.1.2. Content-based Gender Inference

Content-based methods focus more on the content posted by Twitter users online. Twitter allows the users to post 140-character tweets(280-character after Nov.7th, 2017 [3]) on their personal account. Early researches have proven that user of different genders have different word choices and writing styles. For example, Rao et al. tries to processes text generated by Twitter user to extract unigram and bigram features using a Support Vector Machine (SVM) algorithm to determine latent user attributes like gender information (Rao et al., 2010). Several other studies have been using similar n-gram features in combination with logistic and linear regression models to infer more demographic information of the user like gender (Burger et al., 2011), age (Nguyen et al., 2013a;b) ,politic attitudes (Pennacchiotti & Popescu, 2011).

In addition to n-grams features, stylistic features in Twitter text have also been used for user gender and other demographic information inference. For instance, several approaches describe methods to determine gender based on the usage of gender specific words like he, she or his, her, abbreviations, punctuation (Fink et al., 2012), smileys, repeated letters, pronouns, EMOJI (Wolf, 2000) , hashtags and other grammatical features (Cheng et al., 2011; Ito et al., 2013).

However, content based methods require background and domain specific knowledge of natural language processing and can not be easily extended to other languages. Some studies have made some progresses on this direction but more efforts are needed (Mozetič et al., 2016).

2.1.3. Hybrid Gender Inference

Hybrid approaches try to combine both profile based methods, content based methods and other source features or information to improve the accuracy of results. Many efforts have been devoted for this methods, for example, Orlandi et al. tries to use information both from other sources like Facebook and Twitter to infer user profile (Orlandi et al., 2012). Li et al. tries to use online social networks to help user identification in Twitter (Li et al., 2017). Other directions like using hierarchical knowledge base (Kapanipathi et al., 2014), Twitter account the user following (Chamberlain et al., 2016) or migration patterns (Zagheni et al., 2014) etc.

In this paper, we want to use the third approach, the hybrid method to infer the user gender information. To be more specific, we want to combine Twitter users' mobility features like travel pattern with the content based information to design a new system to infer Twitter user gender. Previous studies have already shown that there are differences in Twitter user geo-temporal distribution between different genders (Graham et al., 2014; Mahmud et al., 2014; Weber & Garimella, 2014; Longley & Adnan, 2016). But few work has been done to utilize the geo-temporal feature of Twitter accounts to infer the gender information.

2.2. User Temporal and Spatial Travel Patterns

Human trajectories show a high degree of temporal and spatial regularity (Gonzalez et al., 2008). Two important statistical properties of such mobility patterns are displacement length (i.e., distance between a person's positions at consecutive locations) and radius of gyration (i.e., characteristic distance traveled by a person within a specific period of time)(Gonzalez et al., 2008). Besides, three kinds of entropy are calculated to quantify the degree of predictability of a person's travel trajectories (Song et al., 2010). Respectively, they measure location diversity, heterogeneity of visitation patterns along time and the full temporal and spatial order of the person's mobility pattern. Moreover, the number of co-location records are summarized within different time period to measure the size of the set of co-locations in social network studies (Cranshaw et al., 2010).

Although different variables characterizing detailed temporal and spatial travel patterns have been proposed , they are based on dense cell phone record data or verbose sur-

[1] `https://github.com/tue-mdse/genderComputer`

[2] `https://github.com/muatik/genderizer`

[3] `https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html`

vey data. For the rising social media data, temporal and spatial travel patterns are usually visualized for interactive investigations (Yin et al., 2016), while lacking numeric parameters for further statistical analysis. Different from previous dense data study, social media data like Twitter is sparse, which only captures noncontinuous segments of users' travel trajectories. So it is hard to get enough useful features from Twitter data. To use the travel patterns stored in social media data like Twitter, our project tries to design new temporal and spatial features to represent travel patterns and weigh their relative importance to the classifier through several tests.

3. Methodology

3.1. Data Construction

We got 254,698 tweets of 8614 users in the city of St. Louis, MO from 4:12:06 AM Sep. 11 2010 to 5:49:50 AM Jul. 6 2014 (data provided by Prof. Qunying Huang in Department of Geography at UW-Madison. This dataset is for use of a larger project while this course project is only a part of it.) Each tweet consists of following informational fields : tweet id, tweet sent to user, tweet device, tweet create at,content, hash tag, tweet retweet count, tweet zone type, user id, user account, user name, user follower, user tweet count, user location, user profile image, user start time, user fname, user lname, user mname, user ethnicity. Among them, we use content, hash-tag, tweet zone type, user account name, user's first name, user's last name, user's profile image, temporal stamp in this project.

Zone type is an important field to indicate different human activities (e.g., stay at home, work and entertainment). There are six zone types in total, linked with ten urban land use types. Details are listed below,

Zone type index 1 Residential : Neighborhood Preservation Area

Zone type index 2 Education, Health, Shopping, Eating, Entertainment, Service:Regional Commercial Area, Recreational/Open Space Area

Zone type index 3 Office : Business/Industrial Development Area, Business/Industrial Preservation Area

Zone type index 4 Transportation, Service : Neighborhood Development Area

Zone type index 5 Office, Entertainment, Shopping, Eating, Service : Opportunity Area, Institutional Area

Zone type index 6 Home, Service, Education, Shopping, Eating, Entertainment, Health, Office:Neighborhood Commercial Area, Neighborhood Development Area

The zone type information is obtained by projecting each geo-tagged tweet onto the St Louis land use map seen Fig. 1 with QGIS (QGIS, 2015).

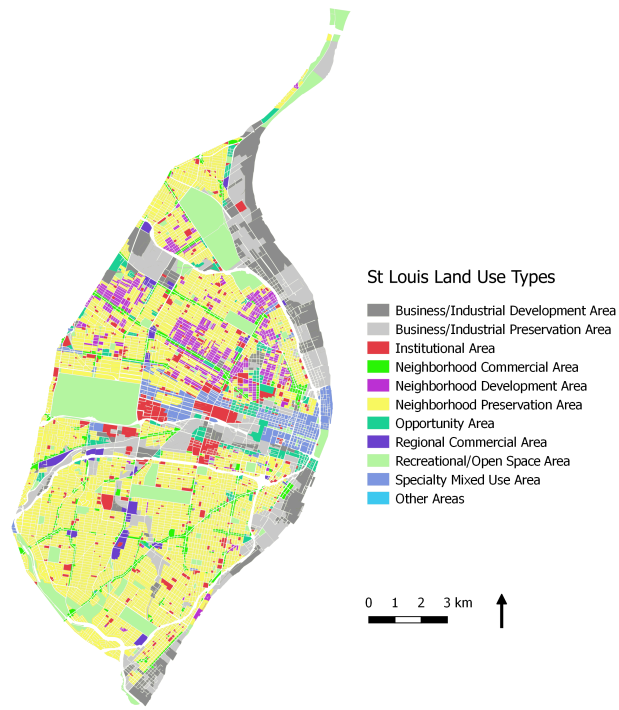

Figure 1. St Louis Land Use Map

If the number of tweets for a single user is too small, travel patterns could not be assessed. Therefore, we removed users with less than 10 tweets, and 237639 tweets of 2522 users were left in our experiment data-set.

3.1.1. Data Annotation

We manually labeled these users with three gender types: male, female and others (including organizations and the uncertain), by investigating their profile images, first/last name and tweet contents.

Usually, we do this task by the following order,

1. Check Twitter user avatar to judge whether user's gender.
2. Check Twitter account's first name.
3. Read tweet message history, to judge from the keywords in tweets like "father", "proud data", "my boyfriend", etc.
4. If still unclear, searching the user's first name, family name in Facebook, Instagram and Google+.

5. If still unclear, check the tweet message interaction between this user and his friends and using tweets like "I am proud of my friends and his girlfriend." to infer the gender.
6. if still unclear, we assign this user gender O which means others or unclear.

3.2. Feature Extraction

In social media like Twitter, a user's tweet contains rich information about: where, when and what. We consider them as spatial features, temporal features and content features for further user gender identification. In this project, we focus on investigating relationships between users' travel patterns, tweet content and their gender affiliation.

3.2.1. Extract Temporal Feature

We selected the following temporal features for each user as candidate temporal features. Details are listed below,

1. Tweet frequency during 12 months (January to December)(12 features)
2. Tweet frequency during weekdays(Monday to Friday) or weekends(Saturday and Sunday).(2 features)
3. Tweet frequency during different time period in one day and we allow overlapping between each time period. And each time period is determined as following using 24 hours(5 features):
   - Morning : 6:00 to 12:00
   - Noon : 11:00 to 14:00
   - Afternoon : 12:00 to 18:00
   - Night : 17:00 to 23:00
   - Late-night : 22:00 to 6:00

Using these features, we try to represent each Twitter users' temporal travel patterns.

3.2.2. Extract Spatial Feature

We selected the following spatial features for each user as candidate spatial features. These features are using land use information to get the frequencies of tweets or the distance the Twitter user travel between each land type. Details are listed below,

1. Frequency of each zone type.(6 features)
2. Travel distance between every two different zone types on weekdays. (15 features)
3. Travel distance between every two different zone types on weekends. (15 features)

Using these features, we try to represent each Twitter users' spatial l travel patterns.

3.3. Extract Content Features

We want to extract content features that can help identify different genders from each Twitter users' text content. Too many features will be used if we directly choose each word in tweet content. We have more than 10,000 unique word in all tweet content, this is overwhelmingly too large compared with spatial features and temporal features. Therefore, we applied naive Bayes network on the words of each user's tweet content and return confidences for different gender types. Naive bayes network is a good model used for word classification due to its swiftness and simpleness (Cheng et al., 2010; Hansen et al., 2011). The network structure is illustrated in Fig. 2.

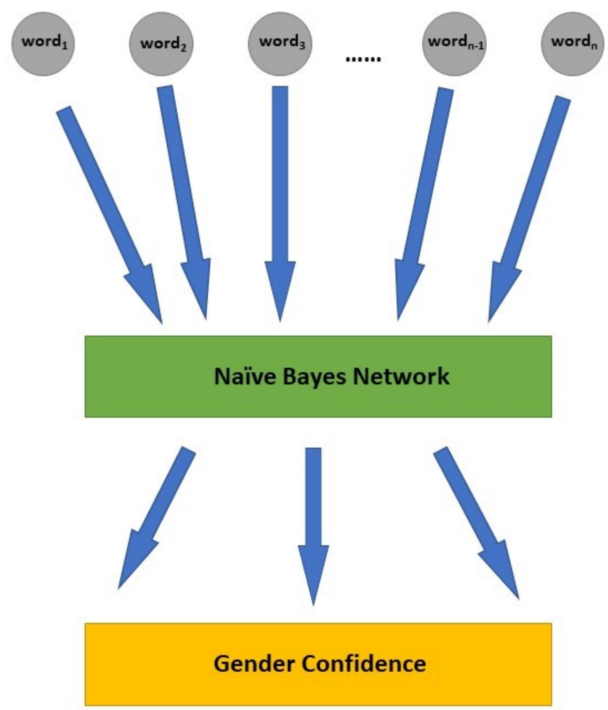

Figure 2. Structure of Naive Bayes Network

Naive Bayes network is based on Bayes Theorem which is shown in Fig. 2. It uses word embedding of each user tweet content to count the frequency of different word types existing in all instances in the train set and predicts the probability of Twitter users' genders, which called gender confidence in Fig. 2.

3.4. Classification

After getting gender confidence from naive Bayes network and combing this feature with 55 spatial and temporal features, we finally decided to use Random Forest to predict the final label of each user' gender. Random Forest classification method was chosen because it could perform classification on high dimensional space by randomized feature selection approach(Breiman, 2001). We have tested different methods based on our data, Random forest outperformed other methods like Neural network, Ada boosting and decision tree which agreed with other studies (Liaw et al., 2002). In short, Random Forest is a form of "ensemble learning" where the algorithm generates a large number of not pruned decision trees and then summarize the results from each decision tree (Breiman, 2001). This means it can provide more accurate predictions with smaller sample sizes and becomes less susceptible to over-fitting and other problems in other classifiers. Therefore, we used the Random Forest functions [4] provided by Scikit-learn and trained the Random Forest classifier with 30 maximum features and 60 allowed sub-trees (Pedregosa et al., 2011).

4. Results

To better test our model and features, we randomly divide the whole data set into development data set(2400 users) and test data set(122 users). For the development stage, And we use 5-fold stratified cross validation to test our model for development stage. Specifically, the 2400 Twitter users are divided into development training set and development testing set. We use development testing set to help choose the best parameters for each model.

After we get the fully developed model, we use the test data set which is not used in development stage to test the accuracy of each model. And to better ensure the accuracy, we do the test of accuracy 5 times for each model. For each test Precision, Recall and F1-score are present.

4.1. Accuracy

First, we present the accuracy result for content-based Naive Bayes Network as shown in Table. 1

We can see that content-based Naive Bayes Network achieve an accuracy of 71 %.

Secondly, we combine gender confidence given by naive Bayes network and travel pattern features to calculate Precision, Recall and F1-score again with different model. All data are averaged 5 times.

[4] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Table 1. Accuracy Table of Content-Based Naive Bayes Network

Experiment	Precision	Recall	F1-score
1	0.67	0.7	0.68
2	0.72	0.73	0.72
3	0.71	0.7	0.71
4	0.72	0.72	0.72
5	0.69	0.69	0.69
Ave.	0.71	0.71	0.71

Table 2. Performance Table of Combination of all feature with different models

Methods	Precision	Recall	F1-score
Random Forest	**0.720**	0.725	0.720
Decision Tree	0.715	0.733	0.723
Extra Tree	0.713	0.718	0.717
AdaBoost Tree	0.718	0.725	0.718
Gradient Boosting Tree	0.715	0.713	0.710
Multi-layer Neural Network	0.688	0.708	0.695

Apparently from Table. 2, we can find that, after combining words and spatial and temporal features, precision, recall and F1-score increases for most models except multi-layer Neural Network which may need more parameter tuning before really achieving a good results.

To further test our method with the benchmark gender inference method using the first name as the vital feature for the inference and the result is shown in Table. 3.

We can see that although our model is not better than benchmark method, the difference is small. And out method does not require that first name must be provided by the user, which is helpful when first name is not provided or the field is not true. And we can see our method can be easily integrated with other methods.

4.2. Spatial and Temporal Feature Importance

As said above, our spatial and temporal feature are easily integrated with other methods, so here we want to show that adding spatial and temporal features can really improve the performance of both profile-based method(first name) and content-based methods. For this part, from the method above, we tried to combine all spatial and temporal features with profile-based method and content-based method and output the relative importance of this feature in Random Forest. We only showed the top 10 important spatial and temporal features in Table. 4 and Table. 5.

As is illustrated above, the confidences derived by first-name-based classifier or content based classifier contribute the most to gender classification. Aside from them, several spatial and temporal features were found to slightly improve the classification performance. They are frequency

Table 3. Performance Table of Random Forest with all features and first name based methods

METHODS	PRECISION	RECALL	F1-SCORE
FIRST NAME BASED	0.898	0.785	0.832
RANDOM FOREST WITH ALL FEATURES	0.8230	0.835	0.830

Table 4. Top 10 important spatial and temporal features for profile-based method(first name)

FEATURE NAME	IMPORTANCE
CONFIDENCE OF GENDER M BY FIRST NAME	0.3395
CONFIDENCE OF GENDER F BY FIRST NAME	0.2683
CONFIDENCE OF GENDER O BY FIRST NAME	0.0154
FREQUENCY IN MAY	0.0195
FREQUENCY OF STAY AT HOME	0.0188
FREQUENCY ON WEEKENDS	0.0172
FREQUENCY AT NIGHT	0.0168
FREQUENCY AT LATE NIGHT	0.0167
FREQUENCY AT AFTERNOON	0.0167
FREQUENCY ON WEEKDAYS	0.0161

Table 5. Top 10 important spatial and temporal features for content-based method

FEATURE NAME	IMPORTANCE
CONFIDENCE OF GENDER M BY CONTENT	0.4395
CONFIDENCE OF GENDER F BY CONTENT	0.2890
CONFIDENCE OF GENDER O BY CONTENT	0.0182
FREQUENCY ON WEEKENDS	0.0132
FREQUENCY AT AFTERNOON	0.0120
FREQUENCY OF STAY AT HOME	0.0119
FREQUENCY AT NIGHT	0.0109
FREQUENCY AT LATE NIGHT	0.1082
FREQUENCY ON WEEKDAYS	0.0094
FREQUENCY OF COMMUTING	0.0091

in May, frequency of stay at home, frequency on weekends, frequency at night, frequency at late night, frequency at afternoon, frequency on weekdays and frequency of commuting.

4.3. Geo-Gender Map

Due to time limit, we do not fully utilize the gender information we get for the Twitter users. However, here we show one usage of the obtained gender information in Fig. 3.

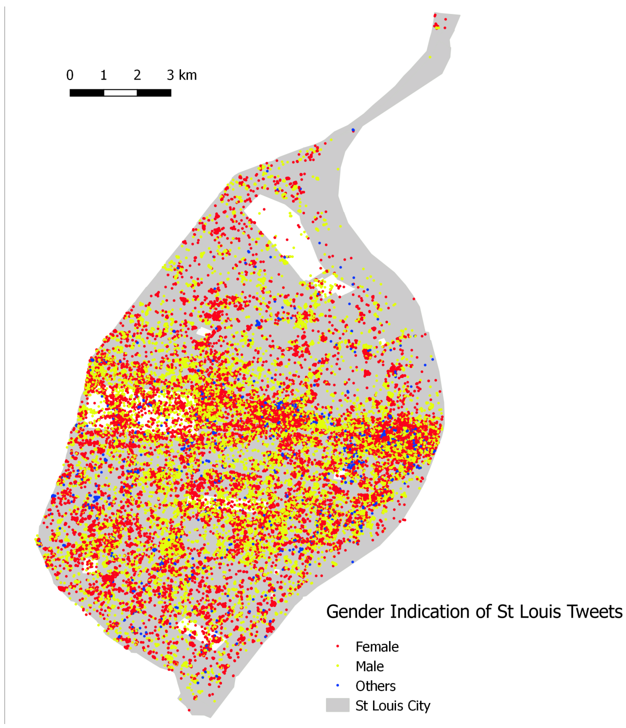

Figure 3. Gender Geo-Distribution of St Louis Tweets

Although it is hard to see any potential usage of Fig. 3, it does show some gender features of St Louis which can be used by local government or police department to help the development of St Louis.

The final Random Forest model is shown in Appendix.

5. Discussion

As shown in Table. 2, with the help of spatial and temporal features to the content mining classifier, precision, recall and F1-score obviously increase under all the models except for Multi-layer Neural Network. We can conclude that travel pattern features works well and contribute positively to gender classification. As for the bad performance on Multi-layer Neural Network, it might result from that we did not acquire proper parameters for it. As shown in table 3, after adding spatial and temporal features to the first-name-based classifier, prediction recall gets clearly improved while the precision decreases. F1-scores of there two methods are nearly the same. It can be said that although our model does not predict as accurate as first name based model, we can get correct prediction within less attempt. And similar F1-scores indicted these two model have similar performance.

Compared with the gender classification based on users' first name, which generally gives a quite good result, our classification method shows a close performance. We cannot obtain users' first name correctly in some situations and it impedes the application of gender classification based on users' first name. However, we can get users' typical words from their contents quite conveniently and give a relatively high accuracy of gender classification with our model.

Word-embedding model is used in naive Bayes network, in which we count the frequency for different word types and construct a word table for that. With the word table we calculate different probabilities for gender types under the condition of these words and give gender confidences. However, one problem is that when classify tweet with the words out of the word table, something has to be done to include these words. To figure this problem we first extend the word table with the new words and then train the naive Bayes network again with the new word table. By using this method our word table becomes larger and larger and the chance to extend the word table becomes smaller and smaller. If the sample data set is large enough we will get a perfect dictionary for content classification. However, this needs more efforts on NLP related area.

6. Conclusions

Content method to process word content gives good results of gender classification of twitter users. Our model combines naive Bayes network and spatial and temporal features to predict Twitter user's gender. It further increases the performance of gender classification. Compared with users' first name gender classification, our model shows a close performance and has less restrictions in application.

In future, we will try to use larger data set with denser sampling of users' geographic locations. Then more accurate spatial and temporal features could be obtained and the performance of our model would be much further improved.

Acknowledgements

We would like to sincerely thank Prof.Liang, Depratment of Computer Science at UW-Madison, for valuable suggestions to implement this project and Prof.Huang, Department of Geography at UW-Madison, for the help of collecting Twitter data-set and annotations of user genders.

Software and Data

We provide all our data and program in Github and you can check them online `https://github.com/iphyer/cs760_TwitterDemographics`.

We use scikit-learn (Pedregosa et al., 2011) as our machine learning program library and Pandas (McKinney, 2015) for data processing.

References

Ahn, Young-joo, Wooten, Dr, Marian, H, Norman, Dr, William, C, McGuire, Dr, et al. Gender differences in travel constraints and changed travel pattern after a senior travel program introduces: A case study of the florence county senior travel program. *Tourism Travel and Research Association: Advancing Tourism Research Globally*, 37, 2016.

Beretta, Valentina, Maccagnola, Daniele, Cribbin, Timothy, and Messina, Enza. An interactive method for inferring demographic attributes in twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pp. 113–122. ACM, 2015.

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

Burger, John D, Henderson, John, Kim, George, and Zarrella, Guido. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309. Association for Computational Linguistics, 2011.

Cesare, Nina, Grant, Christan, and Nsoesie, Elaine O. Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv preprint arXiv:1702.01807*, 2017.

Chamberlain, Benjamin Paul, Humby, Clive, and Deisenroth, Marc Peter. Probabilistic inference of twitter users age based on what they follow. *arXiv preprint arXiv:1601.04621*, 2016.

Chen, Jie, Shaw, Shih-Lung, Yu, Hongbo, Lu, Feng, Chai, Yanwei, and Jia, Qinglei. Exploratory data analysis of activity diary data: a space–time gis approach. *Journal of Transport Geography*, 19(3):394–404, 2011.

Cheng, Na, Chandramouli, Rajarathnam, and Subbalakshmi, KP. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

Cheng, Zhiyuan, Caverlee, James, and Lee, Kyumin. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768. ACM, 2010.

Cranshaw, Justin, Toch, Eran, Hong, Jason, Kittur, Aniket, and Sadeh, Norman. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pp. 119–128. ACM, 2010.

Fink, Clayton, Kopecky, Jonathon, and Morawski, Maksym. Inferring gender from the content of tweets: A region specific example. In *ICWSM*, 2012.

Gonzalez, Marta C, Hidalgo, Cesar A, and Barabasi, Albert-Laszlo. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

Graham, Mark, Hale, Scott A, and Gaffney, Devin. Where in the world are you? geolocation and language identification in twitter. *The Professional Geographer*, 66(4): 568–578, 2014.

Hansen, Lars Kai, Arvidsson, Adam, Nielsen, Finn Årup, Colleoni, Elanor, and Etter, Michael. Good friends, bad news-affect and virality in twitter. *Future information technology*, pp. 34–43, 2011.

Hasan, Samiul, Zhan, Xianyuan, and Ukkusuri, Satish V. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, pp. 6. ACM, 2013.

Ito, Jun, Hoshide, Takahide, Toda, Hiroyuki, Uchiyama, Tadasu, and Nishida, Kyosuke. What is he/she like?: Estimating twitter user attributes from contents and social neighbors. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pp. 1448–1450. IEEE, 2013.

Kang, Chaogui, Gao, Song, Lin, Xing, Xiao, Yu, Yuan, Yihong, Liu, Yu, and Ma, Xiujun. Analyzing and geo-visualizing individual human mobility patterns using mobile call records. In *Geoinformatics, 2010 18th International Conference on*, pp. 1–7. IEEE, 2010.

Kapanipathi, Pavan, Jain, Prateek, Venkataramani, Chitra, and Sheth, Amit. User interests identification on twitter using a hierarchical knowledge base. In *European Semantic Web Conference*, pp. 99–113. Springer, 2014.

Li, Yongjun, Zhang, Zhen, and Peng, You. A solution to tweet-based user identification across online social networks. In *International Conference on Advanced Data Mining and Applications*, pp. 257–269. Springer, 2017.

Liaw, Andy, Wiener, Matthew, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

Liu, Wendy and Ruths, Derek. What's in a name? using first names as features for gender inference in twitter. In *AAAI spring symposium: Analyzing microtext*, volume 13, pp. 01, 2013.

Longley, Paul A and Adnan, Muhammad. Geo-temporal twitter demographics. *International Journal of Geographical Information Science*, 30(2):369–389, 2016.

Mahmud, Jalal, Nichols, Jeffrey, and Drews, Clemens. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):47, 2014.

McKinney, Wes. pandas: a python data analysis library. *see http://pandas. pydata. org*, 2015.

Mislove, Alan, Lehmann, Sune, Ahn, Yong-Yeol, Onnela, Jukka-Pekka, and Rosenquist, J Niels. Understanding the demographics of twitter users. *ICWSM*, 11:5th, 2011.

Mozetič, Igor, Grčar, Miha, and Smailović, Jasmina. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036, 2016.

Nguyen, D, Gravel, R, Trieschnigg, D, and Meder, Theo. "how old do you think i am?": A study of language and age in twitter. In *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, pp. 439–448, 01 2013a.

Nguyen, Dong, Gravel, Rilana, Trieschnigg, Dolf, Meder, Theo, and Yeung, C-M Au. Tweetgenie: automatic age prediction from tweets. *ACM SIGWEB Newsletter*, 4(4), 2013b.

Orlandi, Fabrizio, Breslin, John, and Passant, Alexandre. Aggregated, interoperable and multi-domain user profiles for the social web. In *Proceedings of the 8th International Conference on Semantic Systems*, pp. 41–48. ACM, 2012.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.

Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pennacchiotti, Marco and Popescu, Ana-Maria. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 430–438. ACM, 2011.

Preoţiuc-Pietro, Daniel and Cohn, Trevor. Mining user behaviours: a study of check-in patterns in location based social networks. In *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 306–315. ACM, 2013.

QGIS, DT. Qgis geographic information system. open source geospatial foundation project, 2015.

Rao, Delip, Yarowsky, David, Shreevats, Abhishek, and Gupta, Manaswi. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pp. 37–44. ACM, 2010.

Rao, Delip, Paul, Michael J, Fink, Clayton, Yarowsky, David, Oates, Timothy, and Coppersmith, Glen. Hierarchical bayesian models for latent attribute detection in social media. *ICWSM*, 11:598–601, 2011.

Siła-Nowicka, Katarzyna, Vandrol, Jan, Oshan, Taylor, Long, Jed A, Demšar, Urška, and Fotheringham, A Stewart. Analysis of human mobility patterns from gps trajectories and contextual information. *International Journal of Geographical Information Science*, 30 (5):881–906, 2016.

Sloan, Luke, Morgan, Jeffrey, Housley, William, Williams, Matthew, Edwards, Adam, Burnap, Pete, and Rana, Omer. Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological research online*, 18(3):7, 2013.

Sloan, Luke, Morgan, Jeffrey, Burnap, Pete, and Williams, Matthew. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS one*, 10(3):e0115545, 2015.

Song, Chaoming, Qu, Zehui, Blumm, Nicholas, and Barabási, Albert-László. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

Weber, Ingmar and Garimella, Venkata Rama Kiran. Visualizing user-defined, discriminative geo-temporal twitter activity. In *ICWSM*, 2014.

Wolf, Alecia. Emotional expression online: Gender differences in emoticon use. *CyberPsychology & Behavior*, 3 (5):827–833, 2000.

Yin, Junjun, Gao, Yizhao, Du, Zhenhong, and Wang, Shaowen. Exploring multi-scale spatiotemporal twitter user mobility patterns with a visual-analytics approach. *ISPRS International Journal of Geo-Information*, 5(10): 187, 2016.

Zagheni, Emilio, Garimella, Venkata Rama Kiran, Weber, Ingmar, et al. Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, pp. 439–444. ACM, 2014.