IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 17, 2020, accepted June 2, 2020, date of publication June 8, 2020, date of current version June 16, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3000609*

# Modeling Weather-Aware Prediction of User Activities and Future Visits

**SAMIA NAWSHIN[1], MD. SADDAM HOSSAIN MUKTA[2], MOHAMMED EUNUS ALI[1], AND A. K. M. NAJMUL ISLAM[3]**

[1]Department of Computer Science and Engineering (CSE), Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh
[2]Department of Computer Science and Engineering (CSE), United International University, Dhaka 1212, Bangladesh
[3]LUT School of Engineering Science, LUT University, 53850 Lappeenranta, Finland

Corresponding author: A. K. M. Najmul Islam (najisl@utu.fi)

**ABSTRACT** In recent years, *Location-Based Social Networking* (LBSN) sites such as *Foursquare*, *Facebook Places*, and *Twitter* have become extremely popular due to the extensive usage of location-enabled smart phone technologies. These LBSNs allow users to post their *check-ins* that provide important set of information about users' activities and preferences. Several existing research works predict users' activities from social media *check-ins* data considering various aspects such as *time*, *venue*, and *occurrences*. However, none of the earlier studies investigate the influence of weather on users' preferences of *activities* and *mode of transportation preferences*. Psychological studies show that weather has a strong influence on human activities. In this paper, we predict users' *travel mode*, *day/night time activities* and *future visit* from weather condition derived from social media *check-ins*. In particular, we develop several classification models to predict users' preferable mode of transportation, day/night activities, and future visits from the users' *check-ins* based on different weather conditions. We use two real datasets: *Tokyo* and *New York city* to validate our models. Our classifiers achieve substantial strength (on an average AUC of 72.77%) to predict users' *mode of transportation, day/night activities* and their *future visit preferences* for Tokyo dataset. We also compare performance of the classifiers developed by these two datasets.

**INDEX TERMS** Foursquare, check-ins, weather, correlation, classification.

## I. INTRODUCTION

In current times, social media sites such as *Foursquare, Flicker, Facebook Places,* and *Twitter* have become popular with the proliferation of smart phones and location aware technologies. Users share information of their visiting *venues* (i.e., park, restaurant, and beach), *locations* (i.e., latitude, longitude, and name), and *events* (i.e., workshop, seminar, etc.) in these social media sites by using *check-ins* features. *Check-ins* provide important set of information about users' activities and preferences. Therefore, we can derive new knowledge about users' personalized preferences by analyzing these *check-ins* of different places, venue types, and visiting times. In this paper, we predict users' *mode of transportation, day/night activities,* and *future visit preferences* from weather condition derived from *check-ins* of the social media platform, i.e., *Foursquare*.

According to the social media statistics, Facebook[1] and Twitter[2] have monthly active users of 2,320 and 326 millions, respectively. Posts (i.e., *status updates, tweets, photos,* etc) of these sites are interesting source of research [1]–[4] as these information help to identify users' explicit and implicit behaviors. Users expose a diverse set of human attributes in social media during interactions. Besides, users share location information via *check-ins* by using GPS features of their smart mobile phones in location-based social networking (LBSN) sites. Therefore, we can extract interesting insights (i.e., *users' preferences* and *habits*) by analyzing these diverse and massive amount of social media *check-in* data. For example, researchers identify users' daily routine [5], life-style pattern [6], urban activity pattern [7], and their mobility patterns [8]. Researchers also conduct behavioral studies by using *geo-location* data such as future activities [2], [9], [10] and finding socially relevant venues

The associate editor coordinating the review of this manuscript and approving it for publication was Hongjun Su.

[1] https://newsroom.fb.com/company-info/
[2] https://bit.ly/2UgVo8R

of a city [11]–[13]. Though several studies have been conducted on users' activity analysis by investigating users' *check-in* data, till date no work has been done that inspects the influence of weather condition on users' behavior and activities. Studies on environmental and meteorological psychology [14] describe the importance of weather in relation to human decision making. People may experience the best at a specific place in a certain weather condition [15]. For example, an individual may prefer to visit a sea beach in a sunny weather, whereas she may prefer for an indoor entertainment in a rainy weather. Similarly, a person may prefer to go to coffee shops frequently on cold weather in compare to hot weather. A number of studies [16]–[19] also show that weather condition has strong effect on users' choices and decision making process. However, most prior literature employ only a limited number of weather related parameters to predict human behaviors [20], [21]. Therefore, those models often suffer from low level of prediction capabilities.

Weather influences how entertaining an experience is and therefore peoples' satisfaction is likely to depend on weather condition. Thus, identifying preferences of places and activities based on weather conditions has many potential real life applications. For example, marketers can decide their policies to attract customers and promote their products. Travel agents can predict the future trend of people's visiting area and they can offer different packages to attract customers' attention. Government and policy makers can also take steps regarding preferable transport mode on a given weather condition.

In this paper, we build several classification models to predict users' *preferable mode of transportation*, *Day/Night activities*, and *future visit* from users' *check-ins* and given weather condition. To the best of our knowledge, we are the first to propose such approach for weather aware prediction of users' activities based on social media check-in data. We first collect *Yang's Foursquare datasets* [9] of Tokyo and New York city. Tokyo and New York city have *check-in* instances of 573,703 and 227,428, respectively. Each *check-in* contains information of *mode of transportation* (i.e., Bus, Train, and private transport, etc.), *activity pattern* (i.e., Traveling, Shopping, etc.), and *traveling venues* (i.e., Park, Harbor, etc.). We cross link these *check-in* datasets by using a weather forecast service, *Dark Sky API*.[3] The service extracts weather information with 13 different attributes such as *weather summary, precipitation intensity, wind bearing, wind speed, humidity,* etc. in response to a *latitude* and *longitude* at a specific time. Thus, we collect accurate weather information against each *check-in*. In this way, we create a new dataset that contains *weather* information corresponding to each *check-ins*. Then, we compute correlation between weather information and location categories by using *Chi-Square* ($\chi^2$) [22] and *Fisher's linear discriminant analysis(LDA)* [23] techniques. In this paper, we consider *weather* information as our independent variables and *location* information as our dependent variable. We find significant correlations between weather information and *users' mode of transportation, activities,* and *future visit*. Later, we observe that majority of the class labels of all models suffer from the class imbalance problem. Thus, we remove the class imbalance problem by applying *Synthetic Minority Over-sampling Technique* (SMOTE) [24] re-sampling technique. Finally, we build different classification models to predict users' *preferable mode of transportation, activities,* and *future traveling venues* from weather condition derived from Foursquare *check-ins*. These models obtain an average accuracy (AUC-72.77%) in predicting users' activity/venue/transportation mode from weather parameters.

## II. LITERATURE REVIEW

### A. WEATHER AND ITS IMPACT ON HUMAN BEHAVIOR

We conduct a literature review on the possible connection between weather and users' preferences. These studies mainly cover the broader research areas of meteorological and environmental psychology, transportation and tourism, and human behavior, among others. Weather has impact on a number of dimensions in our real life preferences. Cassidy [20] discusses in his environmental psychology book that weather affects our lives in many ways. Koetse *et al.* [25] show that weather has strong association with trip generation, transport mode, and destination, i.e., *venue*. Warm and dry weather condition influence outdoor leisure, i.e., *visiting beach* and *park*. Rain, snow, windy, cold, and hot weather have impact on selection of transportation mode and decreased number of destination [21], [26], [27]. Tao *et al.* [28] find that changes in particular temperature and rainfall induce significant number of bus riders. Extreme weather (i.e., *heavy precipitation* and *low temperature*) is known to have an impact on the quality of public transport services. Guo *et al.* [29] find that use of public transport (i.e., bus or rail) is negatively influenced by precipitation. Böcker *et al.* [30] find that precipitation may influence people's daily activities. Spinney *et al.* [31] report that negative precipitation influences sport and outdoor activities. Chan *et al.* [14] also find that negative precipitation affects on physical activities. Brandenburg *et al.* [32] find that recreational events and activities such as visiting bar, and night club have strong relation with weather. Weather exposure sometimes dominate to predict individual travel, other outdoor and indoor activities [25]. Though a number of parameters can be linked with individual behavior, combination of weather parameters dominate our daily activities in reality [30]. A few studies show that temperature is one influencing factor with people's behavioral pattern, but integrated weather indices may demonstrate people's behavioral response better [20], [21].

[3] https://darksky.net/dev/

### B. PREDICTING USERS' DAILY ROUTINE AND LIFE-STYLE PATTERNS

The study of human activity patterns from check-ins is gaining attention rapidly. Several studies [5]–[7] have appeared in the literature. These studies largely rely on the continuous tracking of user location.

Pianese *et al.* [5] predict user's daily routine from her check-ins. The authors investigate when and where the user used to take breakfast, lunch, and where does she go every day. Instead of considering growing sensor data, the authors analyze data from multiple social networks. They use automated techniques for filtering, aggregating, and processing social networking traces to extract regular occurring user activities. Geo-location data from social media offers new ways to understand users' preference of interests and actions. In this regard, Hasan *et al.* [6] explore the idea of inferring individual life-style patterns from activity-location choices, revealed in social media. The authors discover the contextual information or location categories of check-ins performed by users. They infer individual geo life-style patterns through building probabilistic topic models.

Location-based social network generated data contains rich information on the whereabouts of urban dwellers. Such data reveals who spends time where, when, and on what type of activity. Çelikten *et al.* [7] make a probabilistic model with minimal assumptions about the data using Foursquare check-in data. They extract many interesting information about urban activity pattern from users' check-ins of visiting locations. These interesting information are about the places or regions of the city, which places are similar to each other in the city, what are the features that distinguish one region from another. Zadeh *et al.* [33] conduct a study on flu outbreak in the US from Twitter location aware dataset. They compute both spatial and temporal analyses and observe that flu related traffic over social media is similar to the actual outbreak.

It is possible to study individuals' mobility patterns at a fine-grained level from social media data. In [8] authors analyze the check-in patterns of users in LBSN. They find that users' mobility pattern is correlated with social interactions. They observe significant temporal clustering within check-in activities. Human mobility exhibits structural regularities though they change over time. The authors include three approaches to describe these check-in dynamics. They find that, (1) users' behavior is strongly influenced by his/her own recent activity, (2) social influence for example, a visit by a user triggers future visits by his friends and (3) exogenous effects, which include external events. In this work the authors are interested in assessing the effect of social influence on visiting patterns of users. They conduct the study using *Gowalla* dataset.

### C. PREDICTING FUTURE ACTIVITIES OR INFERRING ACTIVITY PREFERENCES

In LBSNs, users interact with different points of interest (POIs) by physically being present there in real-time and leaving their comments. These large-scale user generated digital footprints bring an opportunity to understand the spatial and temporal features of user activity. Yang *et al.* [9] propose an approach to predict users' activity preferences by mining the spatial and temporal features of user activities. First, they model the spatial and temporal activity preference separately, and then use a uniform way to combine these data to infer activity preference. Rahimi *et al.* [34] recommend users' preferred location based on their behavior and temporal pattern by using Gowalla dataset.

Chong *et al.* [10] propose another way to predict venues that a user tends to visit based on historical information of his/her other or previously visited venues using Foursquare data. The approach generates a rank that predicts a number of places where a user likes to visit based on his priority. They explore Latent Dirichlet Allocation (LDA) topic models for venue prediction. Huang [2] introduces a new methodology to predict individual's next location based on sparse footprints accumulated over a long time period by using Twitter data. Laniado *et al.* [35] find an association between *geographical distance* and *social tie* among social media friends. The authors analyze the relationship among 10 million active users of *Tuenti*[4] social media site. Wu *et al.* [36] propose a new dynamic model to predict user's evolving behavior from social media interactions.

### D. CLUSTERING SOCIALLY RELEVANT VENUES

Understanding individual and collective mobility patterns is important for many applications. Cho *et al.* [11] examine the similarity of users based on the venues they have visited in the past. They use network structure information to cluster venues so that a venue's group reflects its functionality. Based on the functionality of the venues' group they can find the similarity of the users. Qu *et al.* [12] conduct trade area analysis from user generated mobile location data. The identification of places with similar usage in urban region is an interesting topic for authorities, urban analysts and residents. For example, Rösler *et al.* [13] present an approach to segment city areas into clusters based on users' activities from LBSN's data.

In the light of above discussion, we observe that no study till now presents users' activity based on weather condition. Especially, prior literature did not use a comprehensive list of weather parameters to predict people's preferences. Indeed, Horanont *et al.* [16] explore the effects of weather on people's everyday activity by using GPS traces of mobile phone users and considering 3 weather parameters (temperature, rainfall and wind speed). Though Horanont *et al.* [16] consider weather issue, they do not use social media data. In addition they ignore other important weather parameters. Therefore, to the best of our knowledge, our work is the first study that builds machine learning based classifiers to predict users' preferences from social media data based on a comprehensive list of weather parameters.

[4] https://www.tuenti.es/

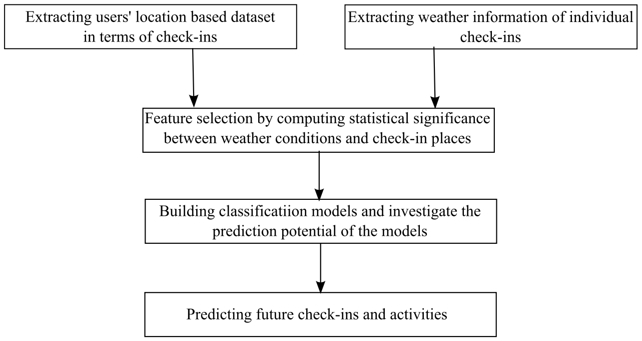
Extracting users' location based dataset in terms of check-ins
Extracting weather information of individual check-ins
Feature selection by computing statistical significance between weather conditions and check-in places
Building classificatiion models and investigate the prediction potential of the models
Predicting future check-ins and activities

**FIGURE 1.** Architecture of our weather aware prediction model.

## III. METHODOLOGY

In this section, we present our approach to develop classification models to predict *users' mode of transportation, visiting places*, and *future activities* from users' check-ins and given weather condition. First, we extract users' location based dataset (i.e., check-ins) and weather information of those check-ins. Then, we select features for our model by computing various statistical significance tests between *weather conditions* and *check-in* places. Next, we build the classification models based on the selected features and investigate the prediction potential of our classification models. Finally, we predict transport mode, day/time activities and future visits of users by using our models.

In this paper, we largely follow the similar approaches for building all models. For example, we use two different statistical techniques for *categorical* and *numeric* feature selections for all models. To avoid repetitive content, we explain elaborately these approaches in this section. The whole process can be summarized in the following steps.

i. **Extracting location based dataset.** We collect *Yang's Foursquare datasets* [9] of Tokyo and New York city. Both datasets contain check-ins of individual users of different venues of the cities. Each check-in data is associated with its *time stamp, its GPS coordinates* and *its semantic meaning* (represented by fine-grained venue-categories).

ii. **Extracting weather information.** We extract weather information of every single *check-in* from a weather forecast service, *Dark Sky API*. In the location based dataset, every check-in data contains the *latitude* and *longitude* of every checked-in location along with the *timestamp*. *Dark Sky API* provides the weather information of that particular location and time.

iii. **Data Pre-processing.** In our dataset there are two types of variables: i) *weather information* and ii) *check-in* places. We consider *weather* information as independent variables and *check-in* places as dependent variable. The attribute, weather summary has 38 different values. For building accurate classification models, we reasonably narrow down the 38 weather categories into 6 broader categories.

iv. **Feature Selection.** To build potential classification models, we select relevant features. For feature selection, we perform different statistical significance tests between weather attributes i.e., *temperature, humidity, wind speed, condition summary,* etc. and location categories from *check-ins*. In our dataset, we have 13 different independent variables related to weather. The weather attributes are *weather summary, weather icon, precipitation intensity, precipitation probability, temperature, apparent temperature, dew point, humidity, wind speed, wind bearing, visibility, cloud coverage* and *air pressure*. Among the weather attributes, *weather summary* and *weather icon* are categorical variables while the rest of the 11 attributes are continuous variables. For feature selection, we follow two different approaches. Since our independent variable is a *mixed*, having both *categorical* and *numerical* attributes, we follow multiple statistical significance tests for feature selection.

- **Feature selection for categorical weather information.** We conduct *Chi-Square* ($\chi^2$) test [37] to check the correlation between categorical weather attributes (i.e., *summary* and *icon*), and dependent variable (i.e., transportation mode, day/night time activities, etc.).
- **Feature selection for numerical weather information.** Our weather condition has 11 numerical independent variables. They are *precipitation intensity, precipitation probability, temperature, apparent temperature, dew point, humidity, wind speed, wind bearing, visibility, cloud coverage* and *air pressure*. For selecting important predictors, we use *Fisher's linear discriminant analysis (LDA)* [23]. Discriminant analysis finds correlation between independent variables and dependent variable having more than two class labels.

v. **Building classification models.** We apply different classification techniques such as *NaiveBayes [38], RandomTree [39], RandomForest [40] and REPTree [41]* to predict *mode of transportation*, *users' activity, and visiting place* based on given weather attributes. We use WEKA [42] machine learning toolkit to run these classifiers. Then, we select the best classifier for building our model. We calculate the performance of our classifiers by using AUC values under the 10-fold cross validation with 10 iterations.

vi. **Handling Class Imbalance Problem.** We notice that majority of our built models are biased that might predict wrong class label due to the imbalance of class distribution. Therefore, we follow an approach to solve such class imbalance problem by using over-sampling technique. In our experiment, we use *Synthetic Minority Over-sampling Technique* (SMOTE) [24] re-sampling technique that uses a subset of data from the minority class and creates new synthetic similar instances.

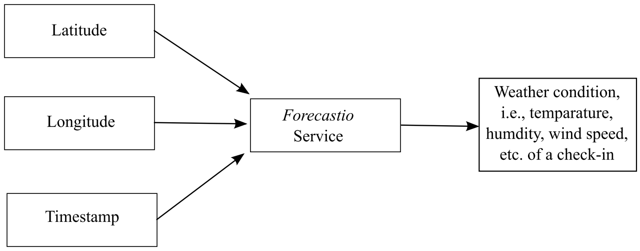
Latitude
Longitude
Timestamp
Forecastio Service
Weather condition, i.e., temparature, humdity, wind speed, etc. of a check-in

**FIGURE 2.** Extracting weather information.

vii. **Re-Building the Classification Model.** We again apply different classifiers by using our dataset with WEKA [42] machine learning toolkit.

## IV. DATASET PREPARATION

As mentioned earlier, we have collected Yang's Foursquare dataset[5] [9] of New York city and Tokyo city. The New York dataset contains 227,428 check-ins of 1,083 individual users and the Tokyo dataset contains 573,703 check-ins of 2,293 individual users of 251 different venues. Each check-in data is associated with *anonymized user id, Foursquare venue id, Foursquare venue category id, Foursquare venue category name, latitude* and *longitude* of the venue, timezone offset in minutes between when the check-in occurred and the same time in coordinated universal time (UTC).

Then we cross-link our check-in dataset with a weather forecast service, *Dark Sky API*,[6] to collect the weather information of every single check-ins. We collect weather information by feeding latitude, longitude, and timestamp of every single check-ins by using the API. The API provides *summary, weather icon, precipitation intensity, precipitation probability, temperature, apparent temperature, dew point, humidity, wind speed, wind bearing, visibility, cloud coverage* and *air pressure*. Figure 2 briefly presents the weather information extraction process.

To get data from the weather forecast service, *Dark Sky API*, we use python-forecast.io 1.4.0, a thin python wrapper.[7] The *dark Sky API* provides us the weather information of anywhere on the globe. The API serves two types of requests: i) *weather forecast for the next week*, and ii) *weather conditions for a date in the past*. We use the second type of API request to get the weather information of all the check-in data points. The API offers various kind of weather information in 39 different languages including apparent (*feels-like*[8]) temperature, atmospheric pressure, cloud coverage, dew point, humidity, liquid precipitation rate, moon phase, nearest storm distance, nearest storm direction, ozone, precipitation type, snowfall, sunrise and sunset, temperature, text summaries, wind gust, wind speed and wind direction. The API requests return weather condition in a JSON format. Then we parse the response directly and collect the required weather information. Later, we write these information in a *.csv* file. We find several studies [43]–[45] that use *forecast.io* Python wrapper for collecting weather information in their research. Zhang *et al.* [46] also build a location aware dataset for their study by making a fusion from different sources.

**TABLE 1.** Initial dataset and the dataset after applying SMOTE of Tokyo for preferable transport mode.

| Class labels | Initial Dataset | | After applying SMOTE | |
|---|---|---|---|---|
| | # of instances | (%) | # of instances | (%) |
| Train | 199,456 | 78.4% | 199,456 | 63% |
| Subway | 41,460 | 16.3% | 41,460 | 13% |
| Light Rail | 2,972 | 1.17% | 23,776 | 7.5% |
| Bus | 7,930 | 3.12% | 31,720 | 10% |
| Private Transport | 2,516 | 1% | 20,128 | 6.4% |
| Total | 254,335 | 100% | 316,540 | 100% |

## V. BUILDING MODELS

In our study, we build four different models to predict user's mode of preferred transportation, day activity, night activity, and visiting places from different weather conditions. We use *Tokyo* dataset as default dataset in this experiment. Later, in Subsection V-E, we briefly explain about the models for New York dataset.

### A. TRANSPORT MODE PREDICTION

Our first model predicts preferable mode of transport on various weather condition. Our model has five different class labels: *Bus, Light rail, Private transport, Subway* and *Train*. Table 1 shows the Tokyo city dataset for building the model. The dataset contains 254,335 instances. The class *Train* is the majority class having 78.4% of instances, where *Private Transport* is the minority class having only 1% of instances in our dataset.

We select the appropriate features for building model. We follow the feature selection process that we describe in item (iv) of Section III (Methodology). In our dataset, we have 13 different independent weather variables and transport mode is our dependent variable. Our dependent variable, i.e., transport mode, is categorical.

For feature selection of categorical weather information and transport mode, we conduct *Chi-Square* ($\chi^2$) test to check the correlation between *preferable transport mode* and weather attributes, i.e., *summary* and *icon*. We find that users' preferable transport mode and weather conditions are correlated. For the attribute *weather summary*, we find the value $\chi^2 = 81.808$ and the *degrees of freedom* (df) = 20. For the attribute *weather icon*, we find the value $\chi^2 = 166.065$ and the *df* = 32.

For feature selection of numerical weather information and transport mode, we use Fisher's LDA. Our dependent variable, i.e., preferable transport mode, has 5 different class labels. We find that the 3 attributes *temperature, apparent temperature (or feels like)* and *dew point* are highly correlated with each other having correlation coefficients larger

[5] https://sites.google.com/site/yangdingqi/home/foursquare-dataset
[6] https://darksky.net/dev/
[7] https://pypi.org/project/python-forecastio/
[8] https://www.home-assistant.io/integrations/darksky/

**TABLE 2.** Fisher's linear discriminant function coefficients between weather attributes and transport mode.

| Predictors | Bus | light Rail | Private Tran. | Subway | Train |
|---|---|---|---|---|---|
| Prec. inten. | 182.531 | 181.187 | 182.004 | 181.934 | 182.969 |
| Prec. prob. | -19.503 | -19.833 | -19.605 | -19.511 | -19.507 |
| Apparent temp. | 24.165 | 24.308 | 24.268 | 24.165 | 24.306 |
| Humidity | 209.324 | 210.008 | 209.912 | 208.647 | 208.953 |
| Wind speed | 8.261 | 8.369 | 8.261 | 8.236 | 8.249 |
| Wind bearing | 0.282 | 0.283 | 0.282 | 0.283 | 0.282 |
| Visibility | -0.177 | -0.183 | -0.130 | -0.182 | -0.185 |
| Cloud coverage | -0.698 | -0.295 | -0.703 | -0.237 | -0.508 |
| Air pressure | 23.946 | 23.950 | 23.942 | 23.940 | 23.943 |

than 0.8. Therefore, we discard *temperature* and *dew point* attributes from the feature list for our model building.

Table 2 shows Fisher's linear discriminant function coefficients between weather attributes and transport mode. These coefficients are helpful in deciding which variable affects more in classification. From these features, we select 8 numeric attributes: *precipitation intensity, apparent temperature, humidity, wind speed, wind bearing, visibility, cloud coverage* and *air pressure*. As mentioned earlier, we also use 2 categorical variables, *weather summary* and *weather icon*, as predictors.

Next, we apply *Naive Bayes, Random Forest [47], Random Tree [48]* and *RepTree* [41] classifiers in our dataset by using WEKA [42] machine learning toolkit. We calculate the performance of the classifier by using AUC values under 10-fold cross validation. We observe that the performance of Random Forest Tree Ensemble[9] is the best. Therefore, we finally choose Random Forest Tree Ensemble as our classifier. We find that on an average the AUC value of our classifier is 66.6%. We also find that our model has MAE and RMSE values of 0.1279 and 0.2628, respectively.

The classification result shows that the AUC values are moderate. Classifiers tend to predict class label that has large number of training instances. In our built model, we observe that TPR rate of Train class label is strong (0.948) and the rest of the class labels are poor. We investigate that the inconsistency among the performance of different class labels are due to class imbalance problem in our dataset. From Table 1, we find that *Train* and *Subway* class labels have 78.4%, and 16.3%, respectively (see initial dataset). Instances of the rest of the 3 classes are low.

In our experiment, we use *Synthetic Minority Over-sampling Technique* (SMOTE) [24] re-sampling technique that uses a subset of data from the minority class and creates new synthetic similar instances. SMOTE avoids to make exact replicas of minority class instances to overcome the over-fitting problems. The technique is powerful and widely used in high dimensional imbalanced dataset [49]–[52]. We increase the number of instances up to 6 times than previous for 3 minority classes of our dataset. For example, previously the class Private Transport had 1% of instances of the full dataset. We up-sample this class 6 times and finally we get this class having 6.4% of instances of the total dataset. Table 1 shows the class distribution we find after applying SMOTE (follow last two columns).

**TABLE 3.** Classification results for preferable transport mode after applying SMOTE.

| Class | TPR | FPR | AUC |
|---|---|---|---|
| Train | 0.933 | 0.448 | 0.834 |
| Subway | 0.173 | 0.037 | 0.740 |
| Light Rail | 0.804 | 0.006 | 0.974 |
| Bus Station | 0.569 | 0.012 | 0.929 |
| Private Transport | 0.862 | 0.004 | 0.985 |

**TABLE 4.** Dataset of Tokyo for building the classification model for preferable day-time activity.

| Class | # of instances | Percentage |
|---|---|---|
| Traveling | 6683 | 30.7% |
| Shopping at mall | 7069 | 32.5% |
| Watching movie in theater | 5130 | 33.6% |
| Staying at home | 2873 | 13.2% |

We again apply Random Forest classifier with our dataset by using WEKA machine learning toolkit. Table 3 presents the classification results of our newly built model. We find that on an average the AUC value of our classifier is 85.1%. We also find that MAE and RMSE scores of our model are 0.1371 and 0.2581, respectively. In the resmapled dataset, we find an average TPR of 0.67% though *Subway* has low TPR.

### B. DAY-TIME ACTIVITY PREDICTION

Our second model predicts user's preferable *day-time activity* on various weather conditions. *Day-time activity* has four different class labels: *Traveling, Shopping at mall, Watching movie in theater,* and *Staying at home*. Table 4 shows the dataset that we used for building the model. The dataset contains a total of 21,755 instances. Majority of the classes have similar number of instances except the class of *Staying at home*. The class has only 13.2% of the total instances. All other 3 classes have around 30% of instances of the total dataset.

We select the appropriate features for building the classification model. We follow the same process of item (iv) of Section III. We find that weather conditions and users' *day time activity* are correlated. We find that $\chi^2$ values for *weather summary* and *weather icon* are 38.683, and 225.064, respectively. These $\chi^2$ values are statistically significant.

For feature selection for continuous weather information, we use Fisher's LDA to prioritize among the predictors. We find that 3 attributes (i.e., *temperature*, *apparent temperature* and *dew point*) are highly correlated with each other that has correlation coefficient larger than 0.8. We get similar observation from the previous model. We compute LDA 3 times by using each of the 3 predictors along with the other 8 predictors each time. We find that the predictor *apparent temperature* affects most in the classification.

From Table 5, we observe that the attribute cloud coverage has a very low coefficient value, so we discard the variable

[9] https://sebastianraschka.com/Articles/2014_ensemble_classifier.html

**TABLE 5.** Fisher's linear discriminant function coefficients between weather attributes and day-time activity.

| Predictors | Traveling | Shopping at mall | Watching movie in theater | Staying at home |
|---|---|---|---|---|
| prec. inten. | 693.344 | 692.687 | 692.790 | 695.254 |
| prec. prob. | 4.403 | 4.574 | 4.470 | 4.362 |
| Apparent temp. | 92.518 | 92.477 | 92.446 | 92.525 |
| Humidity | 3347.929 | 3345.944 | 3345.350 | 3345.179 |
| Wind speed | 2.542 | 2.535 | 2.528 | 2.569 |
| Wind bearing | 0.263 | 0.263 | 0.264 | 0.262 |
| Visibility | 16.272 | 16.246 | 16.209 | 16.122 |
| Cloud coverage | -79.307 | -79.872 | -79.434 | -79.233 |
| Air pressure | 26.953 | 26.956 | 26.954 | 26.948 |

**TABLE 6.** Classification results of the model preferable day-time activity build using Random Forest algorithm.

| Class | TPR | FPR | AUC |
|---|---|---|---|
| Traveling | 0.589 | 0.273 | 0.654 |
| Shopping at mall | 0.557 | 0.291 | 0.682 |
| Watching movie in theater | 0.469 | 0.169 | 0.662 |
| Staying at home | 0.376 | 0.065 | 0.625 |

**TABLE 7.** Tokyo city dataset for building the classification model for preferable night-time activity.

| Class | # of instances | Percentage |
|---|---|---|
| Visiting nightlife spot | 11647 | 55.9% |
| Staying at home | 9202 | 44.1% |
| Total | 20849 | 100% |

**TABLE 8.** Fisher's linear discriminant function coefficients between weather attributes and night-time activity.

| Predictors | Visiting nightlife spot | Staying at home |
|---|---|---|
| prec. intensity | 857.925 | 859.457 |
| prec. probability | 75.368 | 75.609 |
| Apparent temp. | 81.011 | 80.986 |
| Humidity | 3373.513 | 3374.414 |
| Wind speed | 27.994 | 27.977 |
| Wind bearing | 0.495 | 0.495 |
| Visibility | 14.212 | 14.261 |
| Cloud coverage | 9.283 | 9.312 |
| Air pressure | 29.698 | 29.677 |

**TABLE 9.** Classification results of preferable night-time activity using Random Forest algorithm.

| Class | TPR | FPR | AUC |
|---|---|---|---|
| Visiting night life spot | 0.751 | 0.437 | 0.769 |
| Staying at home | 0.563 | 0.249 | 0.729 |

**TABLE 10.** Tokyo city dataset for building the classification model for preferable visiting places.

| Class | # of instances | Percentage |
|---|---|---|
| Park | 7206 | 30.8% |
| Harbor/Marina | 9051 | 38.7% |
| Indoor Museum | 7077 | 30.3% |
| Sea Beach | 44 | 0.2% |
| Total | 23378 | 100% |

from our feature set. Finally, we select 8 attributes from 11 numeric attributes as features for building our classification model.The selected attributes are *precipitation intensity, precipitation probability, apparent temperature, humidity, wind speed, wind bearing, visibility* and *air pressure*.

To build the classification model, we again apply Naive Bayes, Random Forest, Random Tree and RepTree classifiers in our dataset by using WEKA machine learning toolkit. Table 6 shows the best performance of users' preferable day-time activity by using *Random Forest* classifier. We find that on an average the AUC value of our classifier is 65.1%. We also find that MAE and RMSE scores are 0.3239 and 0.4215, respectively.

## C. NIGHT-TIME ACTIVITY PREDICTION

Our third model is the prediction of user's preferable *night-time activity* based on different weather condition. Our training dataset for night-time activity has two different class labels: *Visiting nightlife spot* and *Staying at home*. Table 7 shows the number of instances of our dataset.

For categorical feature selection, we find that $\chi^2$ values of weather summary and icon are 27.752 and 35.933, respectively. Weather summary and icon have *df* of 5 and 32, respectively. We also find both the attributes *weather summary* and *weather icon* are statistically significant. For feature selection of continuous weather information, we again use Fisher's LDA. We find that same outcome as the previous 2 models, 3 attributes (*temperature*, *apparent temperature* and *dew point*) are highly correlated with each other having correlation coefficients larger than 0.8. Thus, at this point we choose only apparent temperature.

Table 8 shows Fisher's LDA function coefficients between weather attributes and night-time activity. Based on this, we select 8 attributes from 11 numeric attributes as features for building our classification model. The 8 numeric attributes are *precipitation intensity, precipitation probability, apparent temperature, humidity, wind speed, visibility, cloud coverage* and *air pressure*. We also use 2 nominal or categorical variables *weather summary* and *weather icon*.

Table 9 shows that the *Random Forest* classifier performs the best in predicting *night-time activity*. Thus, we choose *Random Forest Tree Ensemble* as our working classifier. We find that on an average the AUC value of our classifier is 72.9%. We also find that MAE of our model is 0.3931. The AUC value shows moderate performance (average AUC is 65.7%).

## D. FUTURE VISIT PREDICTION

Our last model for the Tokyo dataset is the prediction of user's preferable visiting place on different weather conditions. Our model has four different class labels: *Park, Harbor/Marina, Indoor Museum* and *Sea Beach*. Table 10 shows the dataset for building future visit prediction. All the three classes except the class *Sea Beach* have similar number of instances. *Sea Beach* has few instances (0.2%) that makes our dataset imbalanced.

**TABLE 11.** Fisher's linear discriminant function coefficients between weather attributes and visiting place.

| Predictors | Traveling | Shopping at mall | Watching movie in theater | Staying at home |
|---|---|---|---|---|
| prec. intensity | 264.612 | 262.562 | 263.131 | 264.969 |
| prec. probability | 2.194 | 2.449 | 2.461 | 1.904 |
| Apparent temp. | 4.193 | 4.198 | 4.197 | 4.221 |
| Humidity | 211.886 | 210.541 | 209.407 | 210.525 |
| Wind speed | 8.633 | 8.641 | 8.611 | 8.705 |
| Wind bearing | 0.346 | 0.346 | 0.346 | 0.344 |
| Visibility | 20.170 | 20.270 | 20.309 | 20.333 |
| Cloud coverage | -72.028 | -71.734 | -71.251 | -71.755 |
| Air pressure | 23.723 | 23.721 | 23.725 | 23.717 |

We observe that users' preferable visiting place and weather conditions are related. For the attributes, *weather summary* and *weather icon* have $\chi^2$ scores of 26.793 (df = 15), and 723.629 (df = 24), respectively.

Table 11 shows Fisher's linear discriminant function coefficients between weather attributes and visiting place. From this, we select 7 attributes from 11 numeric attributes as features for building our classification model.

To build the classification model, we finally choose 9 independent variables and one dependent variable. Our independent variables are: *weather summary, icon, precipitation intensity, precipitation probability, apparent temparature, humidity, wind speed, visibility,* and *air pressure*. Then we apply *Naive Bayes, Random Forest, Random Tree* and *RepTree* classifiers in our dataset by using WEKA machine learning toolkit. We see that *Random Forest Tree Ensemble* classifier performs the best. Thus, we select the *Random Forest Tree Ensemble* classifier as our working model. Our classifier shows an average AUC and MAE scores are 67.1%, and 0.2871, respectively. The obtained AUC value indicates moderate performance. The TPR rate of all the classes except *Sea Beach* is moderate. According to the Table 10, the *Sea Beach* class is our minority class due to 0.2% of instances among all the class labels. Therefore, we handle the class imbalance problem of our dataset to improve the accuracy of our model.

We again apply *SMOTE* re-sampling technique to remove data imbalance problem. We increase the number of instances of the minority class *Sea Beach* by 3 times. Then, the instances of *Sea Beach* class increases from 0.2% to 1.5% of the total instances. We again apply *Naive Bayes, Random Forest, Random Tree* and *RepTree* classifiers in our dataset by using WEKA machine learning toolkit. Finally, we choose *Random Forest Tree Ensemble* as our working classifier. Table 12 shows the performance of our model. We find that on an average the AUC value of our classifier is 68.0%. We also find that *MAE* score of our model is 0.286. All classes have moderate TPR rate and low FPR rate. The class *Sea Beach* has good correlation with Clear weather. *Indoor Museum* has a correlation with Cloudy weather.

**TABLE 12.** Classification results of preferable visiting place after up-sampling the dataset using SMOTE algorithm.

| Class | TPR | FPR | AUC |
|---|---|---|---|
| Park | 0.387 | 0.228 | 0.626 |
| Harbor/Marina | 0.551 | 0.325 | 0.665 |
| Indoor Museum | 0.527 | 0.203 | 0.739 |
| Sea Beach | 0.781 | 0.001 | 0.980 |

**TABLE 13.** Classification results of the model preferable transport mode by using Random Forest classifier.

| Class | TPR | FPR | AUC |
|---|---|---|---|
| Train | 0.470 | 0.224 | 0.613 |
| Subway | 0.587 | 0.415 | 0.630 |
| Light Rail | 0.307 | 0.011 | 0.718 |
| Bus Station | 0.368 | 0.123 | 0.642 |
| Private Transport | 0.494 | 0.041 | 0.766 |

**TABLE 14.** Classification results of preferable day-time activity model by using Random Forest algorithm on New York city dataset.

| Class | TPR | FPR | AUC |
|---|---|---|---|
| Traveling | 0.331 | 0.195 | 0.606 |
| Shopping at mall | 0.409 | 0.203 | 0.671 |
| Watching movie in theater | 0.411 | 0.134 | 0.723 |
| Staying at home | 0.581 | 0.221 | 0.781 |

### E. PERFORMANCE OF THE MODELS OF NEW YORK DATASET

We apply similar techniques with the New York Dataset that we use with the Tokyo Dataset during experiments. In the following paragraphs, We briefly explain the results only.

#### 1) TRANSPORT MODE PREDICTION

New York city dataset has a total of 23,608 instances. We apply *Naive Bayes, Random Forest, Random Tree* and *RepTree* classifiers on these dataset. We again see that Random Forest Tree Ensemble shows the best performance. Table 13 shows the performance of our model. *Subway* is the majority class having 39.6% instances of total dataset. Unlike to the Tokyo city *Subway* is the main public transport in New York city. Then the second largest class is *Train*. Here, *Light Rail* is the minority class having 3.1% of instances. Other two classes *Bus* and *Private Transport* have 18.9% and 11.2% of instances respectively.

#### 2) DAY-TIME ACTIVITY PREDICTION

We observe that among different classifiers, Random Forest Tree Ensemble shows the best performance. Table 14 shows the performance of our model. On an average the AUC value of the classifier is 69.6%. All the classes have moderate AUC values. Similar to the previous models, we find that the class *Traveling* has good correlation with clear *weather summary*. The class *Staying at home* is more common in rain and snow *weather summary*.

**TABLE 15. Performance of the Random Forest classifier for preferable night-time activity of New York City.**

| Class | TPR | FPR | AUC |
|---|---|---|---|
| Visiting night life spot | 0.900 | 0.790 | 0.652 |
| Staying at home | 0.210 | 0.100 | 0.652 |

**TABLE 16. Classification results of preferable visiting place by using Random Forest algorithm on New York dataset.**

| Class | TPR | FPR | AUC |
|---|---|---|---|
| Park | 0.711 | 0.565 | 0.611 |
| Harbor/Marina | 0.359 | 0.204 | 0.625 |
| Indoor Museum | 0.218 | 0.055 | 0.637 |
| Sea Beach | 0.292 | 0.017 | 0.757 |

### 3) NIGHT-TIME ACTIVITY PREDICTION

We apply Naive Bayes, Random Forest, Random Tree and RepTree classifiers on night-time activity of New York city dataset. Table 15 shows the performance of our classifier (Random Forest).

### 4) FUTURE VISIT PREDICTION

We observe that performance of the model (see Table 16) is not good though the AUC value is showing moderate result of 62.7% on an average. Only the class *Park* has good TPR rate of 71.1%, though its FPR rate is also high (56.5%). All other classes have low TPR rate.

## VI. DISCUSSION

We conduct this study to investigate whether weather has impact on our real life activities. First, we conduct experiments with *Tokyo* dataset, then we also run the same experiments over *New York City* dataset.

For mode of transportation, we observe that majority of the instances are from *Train* class label (78.4%). We also observe that *Train* class label has better accuracy than other class labels. We observe that *weather icon* has greater correlation with *transport mode* selection. People usually consider very high level observation of the weather by seeing *weather icon* only so that they can take quick decision of selecting a transportation mode. From our analysis, we notice that people also tend to select mode of transportation (i.e., Train, Bus, or Light Rail) based on *apparent temperature* (or *feels like*). We again notice from our analysis that *precipitation probability, visibility, wind bearing* (i.e., the direction of the wind comes from), and *cloud coverage* have less effect on people's Transportation selection. Our findings on the higher correlation with the weather icon and lower correlations with other weather parameters are supported by the dual process theory [53], which suggests that humans have two systems for decision making. System 1 is automatic (often unconscious) and consumes less cognitive capacity, while system 2 is the rational or analytic process and therefore, requires more cognitive processing. People have the tendency to use system 1 more often when possible. Our findings imply that people use system 1 to get the high level observation by using the weather icon most of the times. Checking precipitation probability, visibility, and wind bearing requires more cognitive processing. Therefore, people tend to avoid it when possible.

We find that Tokyo dataset shows slightly better performance for *Train, Subway,* and *Private Transports* than the New York city dataset. We justify our findings based on the following reasoning. In a news report [54], Sisson describes some positive issues about Tokyo rail and subway systems. Another report [55] describes that Tokyo is efficient in terms of many aspects (i.e., *Rail* and *Transport* systems) than New York city. Therefore, it may be that the people living in Tokyo rely more on Train and Subway during certain weather conditions.

In our study, we also reveal interesting association between *Day-time* activities and weather condition. We observe from our results that *Humidity, precipitation intensity,* and *apparent temperature* (or *feels like*) are the most influencing factors in users' day-time activities. We observe four different activities that people perform at day time. People usually give less check-ins in their own residence unless there are events such as birthday, anniversary, and cooking special food items, etc. Therefore, we also get less instances of *staying at home*. We find from our results that the class label of *staying at home* has less prediction potential than that of other class labels. We notice that other three class labels for day-time activities: *Traveling, Shopping at mall, and Watching movie in theater* have moderate prediction potential (on an average AUC-66.67%). It is also intuitive that these three day-time activities are strongly linked with weather as these are outdoor activities. For example, in a study [28], researchers find that traveling has direct association with weather conditions. We also observe that *Traveling* and *shopping* day time activities are better predictable with Tokyo city dataset than New York city dataset. On the other hand, we obtain better prediction potential with those day time activities that need less mobility such as *watching movies* and *staying at home* by using New York city dataset.

For our another model, *Night-time activity* has two different class labels: *visiting night life spot* and *staying at home*. We observe from the data that both of the classes are evenly distributed in Tokyo dataset. Thus, we find from our results that our classifiers show strong prediction potential (on an average AUC-74.9%). From correlations of Tokyo dataset, we see that *humidity* is the most influential factor that affects our *Night-time activity*. Humans are susceptible to humidity. Our body attempts to balance its temperature by sweating. If humidity is too high, humans cannot evaporate in the air [56], therefore people feel discomfort. Hence, it makes sense that *humidity* plays an important role in our daily activity. Our finding is in line with the findings of prior studies as well. For example, in a study [57], researchers show that most people prefer to stay in shaded spaces when the humidity is high in Asia. On the other hand, we find moderate prediction potential (on an average AUC-65.2%) for New York city dataset, which is 15% less than that of the Tokyo Dataset.

A study [58] shows that New York city has more average tourists than that of Tokyo city. Since more number of visitors come to New York, it is reasonable to find weak prediction model due to the visitors' diversified behavior depending on their own culture and origin.

Our final prediction model is on future visiting venues. From the data, we see that *Sea Beach* class label has only 0.2% of the total instances. People usually remain busy with water activities when they visit *sea beach* and find less scope to share check-ins. Therefore, the class label has less instances. We find from our analysis that *precipitation intensity* and *humidity* have strong association for future visit prediction. We find that *sea beach* class label is strongly predictable (AUC-98%) from weather information. We observe that Tokyo dataset has better prediction potential (on an average AUC-75.25%) than New York city dataset (on an average AUC-65.75%). We also notice that majority of the future venues have similar prediction potential, but *Sea beach* class label makes a huge difference due to completely different characteristics of the venue. According to a review [55], Tokyo has higher average temperature throughout the year and lower humidity, therefore we may better predict when people visit the *sea beach*.

## VII. CONTRIBUTIONS

The contributions of our work are as follows.

We create a new dataset by using the check-ins of the two popular datasets of Tokyo and New York city containing instances of 573,703 and 227,428, respectively. Then, we cross link each of the *check-ins* with the corresponding *weather information*. We find correlations between users' *activity, visiting place*, and *transportation mode* with a variety of weather parameters to identify the most important attributes that predict people's preferences. Therefore, we contribute to the literature on weather and its impact on human behavior by identifying a comprehensive list of weather parameters that can be used to predict future user activities [20], [21]. We continue to build four different classification models: i) *users' modes of transportation*, ii) *users' activities at day-time*, iii) *users' activities at night-time*, and iv) *users' preferences of visiting places*. Our models obtain an average accuracy (AUC-72.77%) in predicting users' activity/venue/transportation mode from a comprehensive list of weather parameters. We note that while prior literature [20], [21], [26], [27] employ only a few weather parameters (e.g., temperature, precipitation) or weather in general (e.g., rain, snow) for prediction, we have used 13 weather parameters in our experiments. Furthermore, while a prior study [16] has been conducted to identify users' activities from GPS dataset based on weather, our study makes fusion of datasets between social media, i.e., Foursquare and weather. Our study shows better accuracy (avg. AUC-72.77%), in compare to Horanont *et al.*'s [16] regression model that has a weak average prediction of 25%.

## VIII. CONCLUSION

In this paper, we have developed techniques for weather-aware prediction of users' activities from user generated geo-tagged data created by *Foursquare*. We have identified correlation between users' activity and weather condition from user generated social media data. To find correlation, we have created two new datasets containing check-in data and weather information. We have built the datasets by cross-linking two websites: i) Foursquare and ii) Dark Sky. We have created eight different datasets for building four models from the original two datasets of New York and Tokyo city. We have built four different machine learning models: *day and night time activities*, *future vising places*, and *preferable modes of transport* on a given weather condition. These models have shown ranging from moderate to strong prediction potential.

Our models have wide range of real life applications including target marketing, traveling place recommendation and policy making for tourism management. In future, we plan to integrate our model with a real life recommendation system for running tourism business.

## REFERENCES

[1] E. Herder, P. Siehndel, and R. Kawase, "Predicting user locations and trajectories," in *Proc. Int. Conf. User Modeling, Adaptation, Personalization*. Cham, Switzerland: Springer, 2014, pp. 86–97.

[2] Q. Huang, "Mining online footprints to predict user's next location," *Int. J. Geograph. Inf. Sci.*, vol. 31, no. 3, pp. 523–541, 2016.

[3] T. H. Silva, P. O. S. V. de Melo, J. M. Almeida, M. Musolesi, and A. A. F. Loureiro, "You are what you eat (and drink): Identifying cultural boundaries by analyzing food and drink habits in foursquare," in *Proc. 8th Int. AAAI Conf. Weblogs Social Media*, 2014, pp. 1–10.

[4] A. K. M. N. Islam, M. Mäntymäki, and H. Kefi, "Decomposing social networking site regret: A uses and gratifications approach," *Inf. Technol. People*, vol. 33, no. 1, pp. 83–105, 2019.

[5] F. Pianese, X. An, F. Kawsar, and H. Ishizuka, "Discovering and predicting user routines by differential analysis of social network traces," in *Proc. WoWMoM*, 2013, pp. 1–9.

[6] S. Hasan and S. V. Ukkusuri, "Location contexts of user check-ins to model urban geo life-style patterns," *PLoS ONE*, vol. 10, no. 5, 2015, Art. no. e0124819.

[7] E. Çelikten, G. Le Falher, and M. Mathioudakis, "Extracting patterns of urban activity from geotagged social data," 2016, *arXiv:1604.04649*. [Online]. Available: https://arxiv.org/abs/1604.04649

[8] Y.-S. Cho, G. V. Steeg, and A. Galstyan, "Where and why users 'check in,"' in *Proc. AAAI*, 2014, pp. 269–275.

[9] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu, "Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 45, no. 1, pp. 129–142, Jan. 2015.

[10] W.-H. Chong, B.-T. Dai, and E.-P. Lim, "Prediction of venues in foursquare using flipped topic models," in *Proc. Eur. Conf. Inf. Retr.* Cham, Switzerland: Springer, 2015, pp. 623–634.

[11] Y.-S. Cho, G. V. Steeg, and A. Galstyan, "Socially relevant venue clustering from check-in data," in *Proc. KDD Workshop Mining Learn. Graphs*, Chicago, IL, USA, Aug. 2013, pp. 1–6.

[12] Y. Qu and J. Zhang, "Trade area analysis using user generated mobile location data," in *Proc. 22nd Int. Conf. World Wide Web (WWW)*, 2013, pp. 1053–1064.

[13] R. Rösler and T. Liebig, "Using data from location based social networks for urban activity clustering," in *Geographic Information Science at the Heart of Europe*. Cham, Switzerland: Springer, 2013, pp. 55–72.

[14] C. B. Chan, D. A. Ryan, and C. Tudor-Locke, "Relationship between objective measures of physical activity and weather: A longitudinal study," *Int. J. Behav. Nutrition Phys. Activity*, vol. 3, no. 1, p. 21, 2006.

[15] S. Becken, "The importance of climate and weather for tourism: Literature review," Lincoln Univ., Lincoln, New Zealand, Tech. Rep., 2010. [Online]. Available: https://hdl.handle.net/10182/2920

[16] T. Horanont, S. Phithakkitnukoon, T. W. Leong, Y. Sekimoto, and R. Shibasaki, "Weather effects on the patterns of people's everyday activities: A study using GPS traces of mobile phone users," *PLoS ONE*, vol. 8, no. 12, Dec. 2013, Art. no. e81153.

[17] Z. Spasova, "The effect of weather and its changes on emotional state—Individual characteristics that make us vulnerable," *Adv. Sci. Res.*, vol. 6, no. 1, pp. 281–290, Mar. 2012.

[18] K. E. Trenberth, K. Miller, L. Mearns, and S. Rhodes, *Effects of Changing Climate on Weather and Human Activities*. Sausalito, CA, USA: Univ. Science Books, 2000.

[19] E. Howarth and M. S. Hoffman, "A multidimensional approach to the relationship between mood and weather," *Brit. J. Psychol.*, vol. 75, no. 1, pp. 15–23, Feb. 1984.

[20] T. Cassidy, *Environmental Psychology: Behaviour and Experience in Context*. London, U.K.: Psychology Press, 2013.

[21] M. Cools, E. Moons, L. Creemers, and G. Wets, "Changes in travel behavior in response to weather conditions: Do type of weather and trip purpose matter?" *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2157, no. 1, pp. 22–28, 2010.

[22] A. Satorra and P. M. Bentler, "A scaled difference chi-square test statistic for moment structure analysis," *Psychometrika*, vol. 66, no. 4, pp. 507–514, Dec. 2001.

[23] P. A. Lachenbruch and M. Goldstein, "Discriminant analysis," *Biometrics*, vol. 35, pp. 69–85, Mar. 1979.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[25] M. J. Koetse and P. Rietveld, "The impact of climate change and weather on transport: An overview of empirical findings," *Transp. Res. D, Transp. Environ.*, vol. 14, no. 3, pp. 205–221, May 2009.

[26] A. Bigano, J. M. Hamilton, and R. S. J. Tol, "The impact of climate on holiday destination choice," *Climatic Change*, vol. 76, nos. 3–4, pp. 389–406, Jun. 2006.

[27] H. A. Aaheim and K. E. Hauge, "Impacts of climate change on travel habits: A national assessment based on individual choices," CICERO, Oslo, Norway, CICERO Rep. 2005:07, 2005.

[28] S. Tao, J. Corcoran, F. Rowe, and M. Hickman, "To travel or not to travel: 'Weather' is the question. Modelling the effect of local weather conditions on bus ridership," *Transp. Res. C, Emerg. Technol.*, vol. 86, pp. 147–167, Jan. 2018.

[29] Z. Guo, N. H. M. Wilson, and A. Rahbee, "Impact of weather on transit ridership in Chicago, Illinois," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2034, no. 1, pp. 3–10, Jan. 2007.

[30] L. Böcker, M. Dijst, and J. Prillwitz, "Impact of everyday weather on individual daily travel behaviours in perspective: A literature review," *Transp. Rev.*, vol. 33, no. 1, pp. 71–91, Jan. 2013.

[31] J. E. L. Spinney and H. Millward, "Weather impacts on leisure activities in Halifax, Nova Scotia," *Int. J. Biometeorol.*, vol. 55, no. 2, pp. 133–145, Mar. 2011.

[32] C. Brandenburg and A. Arnberger, "The influence of the weather upon recreation activities," in *Proc. 1st Int. Workshop Climate, Tourism Recreation. Int. Soc. Biometeorol., Commission Climate Tourism Recreation*, 2001, pp. 123–132.

[33] A. H. Zadeh, H. M. Zolbanin, R. Sharda, and D. Delen, "Social media for nowcasting flu activity: Spatio-temporal big data analysis," *Inf. Syst. Frontiers*, vol. 21, pp. 743–760, Jan. 2019.

[34] S. M. Rahimi, B. Far, and X. Wang, "Behavior-based location recommendation on location-based social networks," *GeoInformatica*, vol. 23, pp. 1–28, May 2019.

[35] D. Laniado, Y. Volkovich, S. Scellato, C. Mascolo, and A. Kaltenbrunner, "The impact of geographic distance on online social interactions," *Inf. Syst. Frontiers*, vol. 20, no. 6, pp. 1203–1218, Dec. 2018.

[36] R. Wu, G. Luo, Q. Jin, J. Shao, and C.-T. Lu, "Learning evolving user's behaviors on location-based social networks," *GeoInformatica*, vol. 24, pp. 1–31, Mar. 2020.

[37] H. O. Lancaster and E. Seneta, "Chi-square distribution," *Encyclopedia Biostatist.*, vol. 2, 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/0470011815.b2a15018

[38] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, vol. 3, 2001, pp. 41–46.

[39] G. Jagannathan, K. Pillaipakkamnatt, and R. N. Wright, "A practical differentially private random decision tree classifier," in *Proc. IEEE Int. Conf. Data Mining Workshops*, Dec. 2009, pp. 114–121.

[40] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, Jan. 2005.

[41] S. K. Jayanthi and S. Sasikala, "Reptree classifier for identifying link spam in Web search engines," *ICTACT J. Soft Comput.*, vol. 3, no. 2, pp. 498–505, Jan. 2013.

[42] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explor. Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.

[43] C. Trattner, A. Oberegger, L. Eberhard, D. Parra, and L. B. Marinho "Understanding the impact of weather for POI recommendations," in *Proc. RecTour@ RecSys*, 2016, pp. 16–23.

[44] J. Pang, P. Zablotskaia, and Y. Zhang, "On impact of weather on human mobility in cities," in *Proc. Int. Conf. Web Inf. Syst. Eng.* Cham, Switzerland: Springer, 2016, pp. 247–256.

[45] D. Ventura, R. Verborgh, V. Catania, and E. Mannens, "Autonomous composition and execution of REST APIs for smart sensors," in *Proc. SSN-TC/OrdRing@ ISWC*, 2015, pp. 13–24.

[46] B. Zhang, G. Trajcevski, and L. Liu, "Towards fusing uncertain location data from heterogeneous sources," *GeoInformatica*, vol. 20, no. 2, pp. 179–212, Apr. 2016.

[47] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[48] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[49] R. Yang, C. Zhang, L. Zhang, and R. Gao, "A two-step feature selection method to predict cancerlectins by multiview features and synthetic minority oversampling technique," *BioMed Res. Int.*, vol. 2018, pp. 1–10, Feb. 2018.

[50] Y.-D. Zhang, Y. Zhang, P. Phillips, Z. Dong, and S. Wang, "Synthetic minority oversampling technique and fractal dimension for identifying multiple sclerosis," *Fractals*, vol. 25, no. 4, Aug. 2017, Art. no. 1740010.

[51] Y.-D. Zhang, G. Zhao, J. Sun, X. Wu, Z.-H. Wang, H.-M. Liu, V. V. Govindaraj, T. Zhan, and J. Li, "Smart pathological brain detection by synthetic minority oversampling technique, extreme learning machine, and Jaya algorithm," *Multimedia Tools Appl.*, vol. 77, pp. 22629–22648, Jul. 2017.

[52] C. Jia and Y. Zuo, "S-SulfPred: A sensitive predictor to capture S-sulfenylation sites based on a resampling one-sided selection undersampling-synthetic minority oversampling technique," *J. Theor. Biol.*, vol. 422, pp. 84–89, Jun. 2017.

[53] S. Vaisey, "Motivation and justification: A dual-process model of culture in action," *Amer. J. Sociol.*, vol. 114, no. 6, pp. 1675–1715, May 2009.

[54] P. Sisson. (2017). *What New York City's Subway System Can Learn From Ones Around the World*. [Online]. Available: https://bit.ly/36P6hX9

[55] R. Halloran, "New Yorkers and Tokyoites compare life in cities," New York Times, New York, NY, USA, Tech. Rep., 2017.

[56] N. Chandler. (2018). *What is Relative Humidity and How Does it Affect How I Feel Outside?* Accessed: Oct. 2, 2019. [Online]. Available: https://bit.ly/34nLGHe

[57] K.-T. Huang, T.-P. Lin, and H.-C. Lien, "Investigating thermal comfort and user behaviors in outdoor spaces: A seasonal and spatial perspective," *Adv. Meteorol.*, vol. 2015, pp. 1–11, May 2015.

[58] A. Millington, "The 19 most visited cities around the world in 2019," Bus. Insider, New York, NY, USA, Tech. Rep., 2019.

**SAMIA NAWSHIN** received the M.Sc. degree from the Bangladesh University of Engineering and Technology (BUET), Bangladesh. Her research interests include artificial intelligence, machine learning, and data mining.

**MD. SADDAM HOSSAIN MUKTA** received the Ph.D. degree from the Data Science and Engineering Research Lab (DataLab), BUET, in 2018. He is current serving as an Assistant Professor with the Department of Computer Science and Engineering (CSE), United International University (UIU), Bangladesh. He has a number of quality publications in both national and international conferences and journals. His research interests include social network analysis and mining, social computing, data mining, and machine learning.

**MOHAMMED EUNUS ALI** received the Ph.D. degree in computer science and software engineering from The University of Melbourne, in 2010. He worked as a Research Fellow and a Visiting Research Scholar with The University of Melbourne, Monash University, and RMIT University. He is currently a Professor with the Department of Computer Science and Engineering (CSE), Bangladesh University of Engineering and Technology (BUET). He is also the Group Leader of the Data Science and Engineering Research Lab (DataLab), Department of Computer Science and Engineering, BUET. His research falls in the intersection of data management and machine learning. His research interests include a wide range of topic in database systems and information management that include spatial databases, practical machine learning, and social media analytics. His research articles have been published in top ranking journals and conferences, such as *The VLDB Journal*, *Information Systems*, WWWJ, DKE, ICDE, CIKM, EDBT, PVLDB, and UbiComp. Dr. Eunus was a recipient of prestigious UGC Award in the year 2012 for his outstanding research contribution.

**A. K. M. NAJMUL ISLAM** received the M.Sc. (Eng.) degree from the Tampere University of Technology, Finland, and the Ph.D. degree in information systems from the University of Turku, Finland. He is currently an Adjunct Professor with Tampere University, Finland. He is also a Scientist with the LUT School of Engineering Science, LUT University, Finland. He also works as the University Research Fellow with the Department of Future Technologies, University of Turku. He has more than 80 publications. His research focuses on human centered computing. His research has been published in top outlets, such as IEEE Access, *Information Systems* journal, the *Journal of Strategic Information Systems*, the *European Journal of Information Systems*, *Technological Forecasting and Social Change*, *Computers in Human Behavior*, the *Journal of Medical Internet Research*, *Internet Research*, *Computers & Education*, *Information Technology & People*, *Telematics & Informatics*, the *Journal of Retailing and Consumer Research*, *Communications of the AIS*, the *Journal of Information Systems Education*, *AIS Transaction on Human–Computer Interaction*, and *Behaviour & Information Technology*, among others.