Knowl Inf Syst
DOI 10.1007/s10115-016-0936-x

REGULAR PAPER

Prediction of places of visit using tweets

Arun Chauhan[1] · **Krishna Kummamuru**[2] · **Durga Toshniwal**[1]

Received: 1 July 2015 / Revised: 21 January 2016 / Accepted: 14 March 2016

Abstract We study the problem of predicting likely places of visit of users using their past tweets. What people write on their microblogs reflects their intent and desire relating to most of their common day interests. Taking this as a strong evidence, we hypothesize that tweets of the person can also be treated as source of strong indicator signals for predicting their places of visits. In this paper, we propose a novel approach for predicting place of visit within a given geospatial range considering the past tweets and the time of visit. These predictions can be used for generating places recommendation or for promotions. In this approach, we analyze use of various features that can be extracted from the historical tweets—for example, personality traits estimated from the past tweets and the actual words mentioned in the tweets. We performed extensive empirical experiments involving, real data derived from twitter timelines of 4600 persons with multi-label classification as predictive model. The performances of proposed approach outperform the four baselines with accuracy reaching 90 % for top five predictions. Based on our experimental study, we come up with general guidelines on building the prediction model in terms of the type of features extracted from historical tweets, window size of historical tweets and on the optimal radius of query around the place of visit at a given time.

Keywords Information systems · Location-based services · Prediction systems · Temporal and geospatial · Twitter

This work is done during the internship of first author in IBM Watson Labs, Bangalore during May 2014–May 2015.

✉ Arun Chauhan
aruntakhur@gmail.com

[1] Computer Science and Engineering Department, Indian Institute of Technology, Roorkee, India

[2] IBM Research, Bangalore, India

1 Introduction

We are living in an era of microblogging, which revolutionize the way people reflect their interests and intents. One of the most popular among those is twitter; at the time of writing this paper there are 302 million users with significant presence in 33 countries.[1] People share their interests and opinions on this microblogging site [6,8,9]. Twitter processes 500 million tweets per day, each of which contain a maximum of 140 characters. The contents of twitter timeline feeds of users are used as indicative measure of their mood and interests, as stated by Golder et al. [15].

Development in wireless communication and location-acquisition technology allows us to create location-based social network (LBSN) like Foursquare, Gowalla. On these networks people can feed their location along with tips related to location in structured format. Recently, these networks are exploited to develop recommender systems especially for suggesting points of interest (POI). Intent of utilizing LBSN feeds attributed upon structured information (check-in, check-out, tips etc.) in social networks. In spite of this valuable information, as of May 2015, the number of users on Foursquare are 55 million,[2] which are very few in comparison with Twitter users (300 millions).

Considering frequently produced voluminous feeds of unstructured text by users on twitter, we propose a novel approach of predicting next place(s) of visit by user using her recent tweets. Here we regarded common perception and hypothesis that there is a strong interrelation between user's intent and their place(s) of visit; however this hypothesis is later justified through machine learning techniques too. As per our knowledge predicting next place(s) of visit using past tweets presented in this article is first of its kind. In this paper, we are only studying the predictive power of tweets on the location of user at given time. It can be used either to create an offer or recommendation. In real system, (1) other information on the user and/or the merchants can be used, (2) a business logic may need to be added to identify recommendations or offers. For example, a bank can also use the customer tweets to send appropriate offers to meet their business objectives. However, user's offer preferences need to be learned over time. For example, some users are very loyal to some merchants that any offer can not make them go to other merchants in the category. The beauty of this approach is not to use any user related personal information (e.g., age, gender), and it is mainly based on unstructured information.

Our approach mainly consists four steps. The first one being finding places that a given user visited in the past from Twitter user timelines.[3] This task is very challenging mainly because of unusual vocabulary and short messages (140 character limit) on twitter timelines. We apply different NLP techniques in finding visited places from past tweets. Other information like location/geo-coordinates and time of tweet is also inferred from structured information associated with tweets using these NLP techniques. The second step constitutes the extraction of various signals from twitter timelines that indicate user's next place of visit. The challenge here is to understand the intent and mood of user at a given time by using vocabulary in past tweets. The mood reflects state of mind and has a vital role in deciding the activity done by user. Also, vocabulary in past tweets can capture latent factors like age, sex, personality trait etc. [34], which again has impact on the person choosing an activity. The third step of our approach involves building models on various sets of features extracted from users' timelines. Last step includes using additional contextual information, in terms of spatial and temporal

[1] https://about.twitter.com/company.

[2] https://foursquare.com/about.

[3] User timeline is the sequence of past tweets blogged by the user on Twitter.

factors, in refining predictions. As time of the day and spatial range has major influence on type of the activity one does, this information is also incorporated in enhancing the accuracy of model.

We perform extensive experimentation using Twitter timelines data. By analyzing tweets for mentions of place of visit and information from Google Places API, we extract ground truth from tweets obtained from over 4600 user timelines. We extract various features from user timeline including textual features, personality traits and temporal features. We build models to predict the categories of places of visit instead of the actual place of visit. The main motivation behind predicting the categories is for personalized promotions and/or recommendations. We do performance analysis of various learning models and associated features. As no existing work address this problem, we define four baselines to understand the performance of proposed approach. We show that the models in the proposed approach consistently perform better than four baselines. These baselines are defined more to highlight the non-obvious nature of the proposed approach. In our experiments, the best model yields about 90 % accuracy for top five predictions when queried within 300 meters radius. We also analyze the performance of the proposed approach on a set of users that are not included in training. We show that proposed approach performs equally well on new users thus addressing cold start problem.

The major contribution of this paper are:

(1) Generation of ground truth data for predicting places of visit using structured and unstructured data in the tweets,
(2) Building generic models for predicting places of visit from tweets, and
(3) Study of various factors contributing to predicting places of visit.

The rest of the paper is organized as follows: Sect. 2 presents our proposed approach in detail for predicting location of interest. Detailed discussion of results and data sets for experiments are given in Sect. 3. In Sect. 4, related work is summarized. Finally, Sect. 5 concludes this paper.

2 Proposed approach

To restate the problem that is considered in this paper, we want to predict the category of the location that a user will most likely to visit in an area, given her twitter handle and time of the visit. There are two points we want to mention before explaining the four steps of proposed approach to the problem. Firstly, we build the generic prediction model that captures the relationship between vocabulary used in past tweets and visited places, without considering user's demographics. The main advantage of this approach is that we do not need individual specific training data.

Secondly, our focus would be the category of the establishment like restaurant, supermarket, pub, gym, rather than the specific establishment (location name). By doing this, both establishment owners and users can get mutual benefits. For example, if proposed system is predicting restaurant and shopping for user in a given spatial proximity, then all the owners of restaurants and shops in that proximity can send the user their available offers for promotions. Along with this, user can also have an option to choose a place according to her own suitable interest.

```
{
 "created": "Tue Apr 29 21:49:46 +0000 2014",
 "content": "u'Fuckin with younghollywooddre  New York is serious with Park
t.co/0hXS4fj768'",
 "geo" : [40.745086550000003, -73.988538259999999] }
,
{
 "created": "Tue Apr 29 18:42:27 +0000 2014",
 "content": "'Hi jayz @ 40/40 Club http://t.co/TJXNmMByKh'",
 "geo" : [40.743114030000001, -73.989175290000006] }
,
{
 "created": "Tue Apr 29 16:52:41 +0000 2014",
 "content": "'About to smash this  @ White Rose System http://t.co/MX3PMAwv
 "geo" : [40.643140350000003, -74.239397030000006] }
,
{
 "created": "Tue Apr 29 16:15:34 +0000 2014",
 "content": "u'Pole studio almost finished  @XYZ  follow now and stay updat
 "geo" : null }
```


Fig. 1 Sample tweet tagged with geo-coordinates and having some location information

Fig. 2 GPA response to a query

2.1 Assign locations to tweets

We want to mention that from here onwards, we use location and visited place interchangeably in this paper. Assigning locations to tweets is a two-step process. In the first step, we filter all those tweets from the timeline which have location information and tagged with geo-coordinates. For location information we identify tweets using regular expression ("I'm at" or "@ "). Figure 1 shows a sample tweet tagged with geo-coordinates and having location information.

In the second step, we use Google Place API (GPA) [16] to get all the places around that tweet geo-coordinates with in a given radius. GPA typically returns about 60 places around the given geo-coordinates with in a given radius (in meters). Figure 2 shows the response from GPA for the sample tweet in Fig. 1. By using simple similarity matching algorithm, we can check the similarity of the place given by GPA and the text in the tweet. We extract three words after the regular expression match in tweet and match these words with place names from GPA. Here we assume that most of the places' names are of three words. The similarity between the GPA place name, s_1, and the words in the tweet, s_2 is computed as the fraction of the number of common word between s_1 and s_2 in the number of words in s_1. If the similarity is more than a threshold α, then we tagged that tweet with the location given by GPA and refer this tweet to as a *location tweet*.

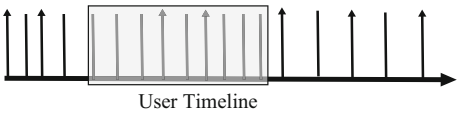

Fig. 3 Location tweets (*upwards arrow*) and usual tweets (non-location tweets) (*vertical line*) on a given user timeline. Location tweets contains phrases like "I am at ABC", where ABC is close to tweets geo-coordinates

At the end of these two steps, we would get the places visited by the user along with the time of visit from the Twitter timelines if appropriate information is present. As mentioned earlier, we consider only the categories of places returned by GPA. You may note that each *location tweet* can be assigned to more than one category, as the corresponding locations could belong to more than one category.

2.2 Extract features

Figure 3 shows graphical representation of a typical user timeline. The tweets shown by arrow ↑ are the *location tweets* and tweets shown by | are the usual tweets (non-location tweets).

We extract various features from the historical tweets by the user, shown as shaded portion on the timeline in the figure, that could help in predicting the future place of visit (i.e., place visited by user just after the shaded window) of user. We consider two sets of features in this paper as follows.

2.2.1 Past tweets for feature extraction

The intuition of using past tweets as feature for training the model is explained as follows: If people are feeling hungry they usually tweet about how they are feeling like "Feeling hungry" or "Mouth watering for Pizza" similarly if they are looking for some outdoor activity their tweet would look like "Boring day!!! Looking for some fun :(". By using these past tweets we can predict the place of visit according to their intent and needs. But the obvious question comes to our mind what could be the length of the window (past tweets) to predict the place of visit for a user. To address this problem we use window length based on time interval. In this approach, we concatenate all tweets in given time window as text document and used it for predicting next place of visit. For example, considering all tweets within 12 h before *location tweet* for predicting place of visit in *location tweet*.

2.2.2 Personality traits as features

We use System U [4] for deriving big five personality traits, fundamental needs and basic human values from the past tweets in a window as explained in Sect. 2.2.1. The big five personality traits include openness, conscientiousness, extraversion, agreeableness, and neuroticism. The profiles from System U contains only numerical value corresponds to these factors in the scale of [0, 1]. Similarly these System U profiles are also used to predict the place of visit for a user. We hypothesize that the personality traits of people could indicate where they visit.

2.3 Building models

As mentioned in Sect. 2.2, we have two types of feature vectors. (a) Tweets from past window labeled by categories of location visited by user just after tweeting these past tweets. As mentioned earlier categories could be more than one. (b) System U profile vectors derived form the past window also labeled by multiple location categories visited by the user. Therefore, we apply two different techniques for each of these feature vectors for building models which are explained as follows.

Window of past tweets We consider past tweets in a given window as a single document (concatenation of tweets with in window) which is used to predict future place of visit. By modeling each document as mixture of underlying topics (categories) where each word is generated from one topic, we can infer future places of visit by using words in tweets. This assumption of modeling the document is very similar to LDA [7] despite topics in our problem are constrained by total number of categories available. This supervised modeling of topic by constraining the number of topics is know as Labeled LDA and was proposed by Ramage et al. [30]. Graphical model of labeled LDA is shown in Fig. 4.

Labeled LDA assumes each label $k \in \{1, \ldots, |\Lambda|\}$ can be described by multinomial distribution β_k over all words in the corpus. The model assumes that each document d uses only subset of labels Λ, denoted by Λ_d, and that document d prefers some labels to others as represented by multinomial distribution θ_d over Λ_d. Each word w in document d is derived from the word distribution of β_z, where $z \in \Lambda_d$. The word is derived using both preferences that is how much the document prefers the label $\theta_{d,z}$ and how much label prefers the word $\beta_{z,w}$. Therefore, from this generative process assumption, an approximate inference algorithm can be used to reconstruct the distribution θ_d over labels, starting from the document itself. For more details on learning and inference in Labeled LDA that we use in this work please refer [30]. Similarly in our case, labels are the categories, and document is the concatenation of past tweets in given window. Therefore, the distribution given by labeled LDA on labels for given test document will give us the confidence for the prediction of each category. Hence, we used labeled LDA for building the models.

We want to mention that when we use SVM and Naive Bayes for our problem (i.e., multi-label classification) as described by [36], we got low accuracy in comparison with labeled LDA. The reason may be that in case of tweets we are having very sparse feature vectors which are used for modeling the SVM and Naive Bayes models. For example, the dimension of feature vector is 700,000 (approx.) because of vocabulary size in corpus but available words in that vector are very few, i.e., approx. 100 (words in a given window). On the other hand, Labeled LDA use less sparse space by using dimensionality reduction technique like topic modeling (Latent Dirichlet Allocation) which gives better accuracy in comparison to SVM and Naive Bayes.

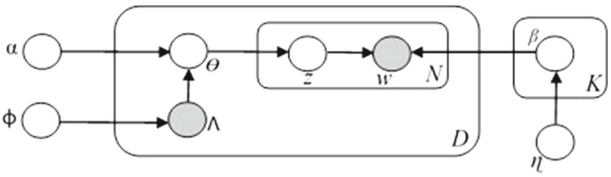

Fig. 4 Graphical model of labeled LDA. α are the parameter of the Dirichlet prior on the per-document topic distributions (θ_d), η are the parameter of the Dirichlet prior on the per-topic word distribution (β_k), ϕ is the vector represents the labels prior (Λ_d) for document d

Input features Input feature in case of labeled LDA is a tuple consist a list of word indices $w = (w_1, w_2, \ldots, w_N)$ and a list of binary presence/absence indicators $\Lambda = (l_1, l_2, \ldots, l_K)$, where each $w_i \in \{1, \ldots, V\}$ and each $l_k \in \{0, 1\}$. Here N is the string (concatenation of tweets in given widow) length, V is the vocabulary size (approx. 700,000) and K (=100) the unique labels in the corpus. l_k in Λ set to 1 for those categories(labels) which are visited by user after a given window and 0 otherwise.

System U profiles We obtain personality traits also referred as System U profiles using System U. System U profile of a user is represented by vector of D dimensions where each dimension contains a value of corresponding trait between 0 and 1. As explained earlier, this System U profile is labeled by visited categories by user which may be more than one. Hence a multi-label classification problem. As vectors are having numerical values we can not use Labeled LDA. Therefore, we use standard method widely know binary relevance method (BRM) [36] used for multi-label classification. BRM transforms the multi-label problem into one or more single label (i.e., binary or multi-class) classification problems. It learns K binary classifiers one for each label present in corpus. It transforms the original dataset into K binary data sets D_j, $j \in \{1, \ldots, K\}$ that contain all examples of original data set, labeled positively if label set of original example contained j and negatively otherwise. Test instance is classified as the union of labels that are positively predicted by K classifiers. BRM also rank the labels based on test instance relevance to particular label. For more details on BRM please see [36]. We have used SVM as a binary classifier for classification in BRM.

Input features For each user, System U profile contains 78 different personality traits having values in [0, 1]. Therefore, feature vector used for training the model have 78 different attributes having values between 0 and 1. This feature vector is labeled by a binary vector $\Lambda = (l_1, l_2, \ldots, l_K)$, where $K = 100$ (total labels in our data set) and each $l_k \in \{0, 1\}$. l_k in Λ set to 1 for those categories(labels) which are visited by user and 0 otherwise.

2.4 Considering additional constraints

To generate final predictions, we consider two additional constraints—radius and time.

2.4.1 Radius of query

Area in which we want to predict is important factor for accurate prediction of user's place of visit, as we do not want to search on all places around the world. Hence, we restrict the place of search. We first use GPA to identify all the places around the given geo-coordinates. We consider the predictions by classification models only within these places identified using GPA. In experiments, we vary radius of search in computing the accuracies of various approaches.

2.4.2 Time of the query

Time is the critical factor in predicting the place of visit for user. As it is very likely for a user to visit some place at particular time. For example, many prefer to visit *park* in the morning and *bar* in the evening. In order to use the time of visit in the prediction, we estimate the probability of visiting a place at a given time from the training data. Let $P(C_i/t)$ represent the estimated probability of any person visiting place C_i at a given time t. Let $P(C_i/Win)$ represent the probability of visiting place C_i by user using her past tweets denoted by Win. In other words $P(C_i/Win)$ is the posterior distribution obtained from the models mentioned

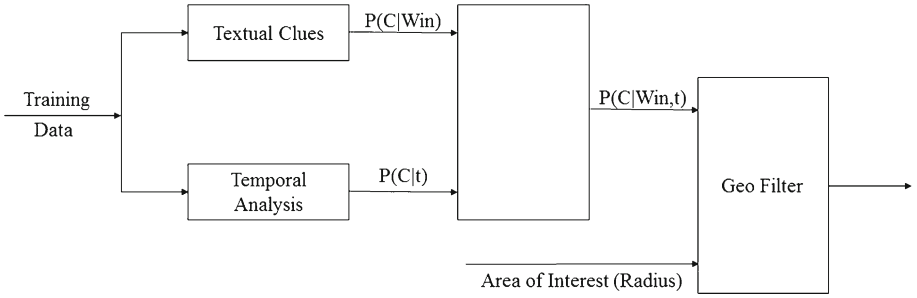

Fig. 5 Block diagram of proposed system

in Sect. 2.3. We combine these two factors under the assumption of independence. We refer the combined classifier as *P(C/Win,t)* for the sake of brevity.

$$P(C_i/Win,t) = \frac{P(Win,t/C_i)P(C_i)}{P(Win,t)} \tag{1}$$

Assuming Win and t are independent, we can write the Eq. (1) as follows:

$$P(C_i/Win,t) = \frac{P(Win/C_i)P(t/C_i)P(C_i)}{P(Win)P(t)} \tag{2}$$

$$= \frac{\frac{P(C_i/Win)P(Win)}{P(C_i)}\frac{P(C_i/t)P(t)}{P(C_i)}P(C_i)}{P(Win)P(t)} \tag{3}$$

after simplifying Eq. (3), we have the following equation

$$P(C_i/Win,t) = \frac{P(C_i/Win)P(C_i/t)}{P(C_i)} \tag{4}$$

where, $P(C_i)$ represent the prior probability of visiting category C_i by any user. For more detail on exact computation of $P(C_i)$ and $P(C_i/t)$ from the training, please refer to Eqs. 5 and 6 respectively.

Block diagram of proposed system is shown in Fig. 5. Using textual clues in past tweets as explained in Sect. 2.2, we obtain probability $P(C/Win)$ by models mentioned in Sect. 2.3. Along with textual clues, probability $P(C/t)$ obtained by temporal analysis is also used for refining the predictions. Finally after combining both probabilities $P(C/Win)$ and $P(C/t)$ under independent assumption as explained in Eq. 4, we get $P(C_i/Win,t)$. Now to generate predictions in area of interest, we use method mentioned in Sect. 2.4.1.

3 Experiments

3.1 Data set

3.1.1 User timelines

We crawled Twitter data which is publicly available by using Twitter search API [37]. We have randomly chosen 4606 users of Twitter whose average tweet rate is at least 20 tweets per day and who have at least tweeted once from New York City between April 24, 2014 and

April 29, 2014. We have collected Twitter timelines of these users containing about 3200 recent tweets.

3.1.2 Identification of location tweets

Among the tweets in the timelines, we identify *location tweets* using the method described in Sect. 2.1. We have used different value of α between interval [0.66, 1]. We found that when $\alpha = 1$, tagging accuracy is approximately 100 %. For evaluation we have taken 100 users' timelines and evaluated it manually. For $\alpha = 0.75$ tagging accuracy is approximately 98 % and for $\alpha = 0.66$ the accuracy is more than 80 %. For our experiments we have used $\alpha = 0.75$.

The distributions of locations associated with the location tweets in the data set is shown in Fig. 6. Even though we have started with the people tweeted from NYC, you may observe that these people have traveled all over the globe.

Among 4606 users, the number of location tweets available on timelines is highly variable. The distribution of the number of location tweets on users timeline is shown in Fig. 7. Please note that each of these location tweets could act as a data sample in our experiments.

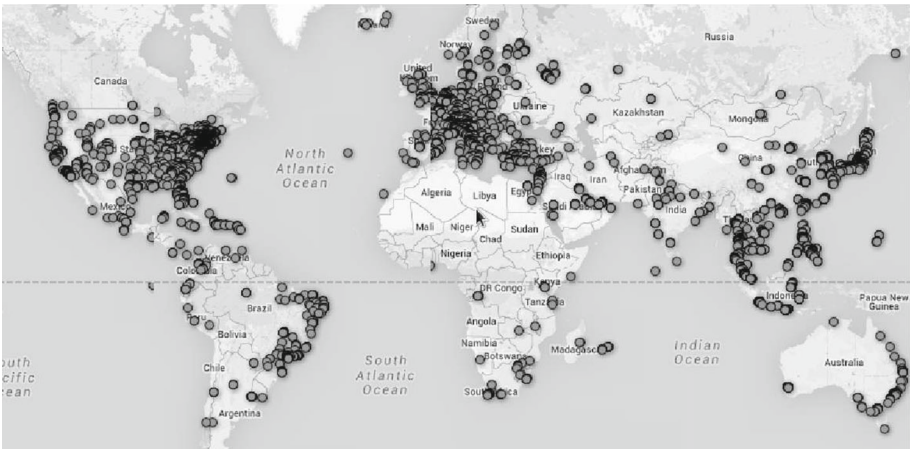

Fig. 6 Distribution of locations associated with the location tweets in the data set

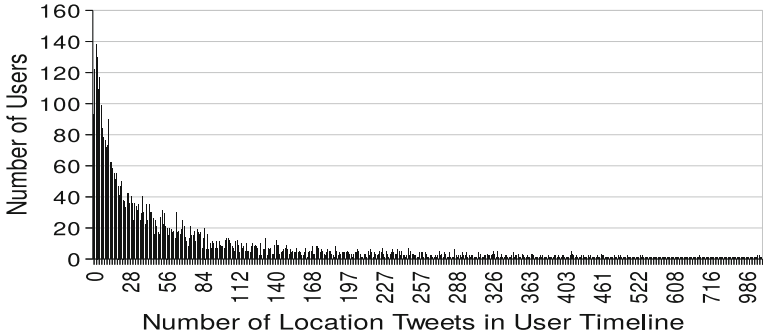

Fig. 7 Number of users versus the number location tweets in their Twitter timeline for all 4606 users

Table 1 Division of users according to number of location tweets on their timelines

Data set	Number of users	Number of location
Data set 1	1706	>60
Data set 2	2900	≤ 60

Table 2 Data sets for training and testing

Data set	# Samples	Detail
Training data set	260,423	Oldest $n-50$ location of each user from data set 1
Testing data set 1	84,470	Latest 50 location of each user from data set 1
Testing data set 2	56,774	All location of each user from data set 2

3.1.3 Training and test sets

For the sake of evaluation, we divided the users into two subsets as shown in Table 1.

(1) *Data set 1* In this set, we considered only those users who are having more than 60 location tweets present on their timelines. Total number of users in this set are 1706.
(2) *Data set 2* In this set we considered the remaining users viz., 2900 users.

Training data set We derive the training data from Data Set 1. We form training data set by using oldest $n-50$ location tweets for creating feature vectors, where n is the total number of location tweets available on users' timeline. Tweets are ordered according to the time when they were generated.

Testing data set We have used two different test data sets named as Test Data Set 1 and Test Data Set 2 in our experiments.

(1) *Test data set 1* We considered only latest 50 location tweets from each user's timeline from Data Set 1 for constructing feature vectors. Then these feature vectors are used to predict the location.
(2) *Test data set 2* We considered all location tweets available on each user's timeline in Data Set 2. This data set is used to evaluate the performance of models on users not seen during the training phase.

These data sets are summarized in Table 2.

3.1.4 Insights into data sets

The total number of location categories we found in our data is 100 which are as follows:

casino, hindu_temple, night_club, subway_station, storage, restaurant, pharmacy, university, aquarium, plumber, taxi_stand, post_office, park, car_wash, painter, general_contractor, lodging, health, home_goods_store, roofing_contractor, place_of_worship, city_hall, bank, transit_station, bar, store, local_government_office, meal_delivery, travel_agency, finance, art_gallery, accounting, spa, lawyer, hair_care, car_dealer, hospital, school, veterinary_care, book_store, hardware_store, physiotherapist, embassy, police, car_repair, moving_company, funeral_home, movie_rental, dentist, museum, cemetery, liquor_store, stadium, bus_station, bicycle_store, insurance_agency, beauty_salon, courthouse, bakery,

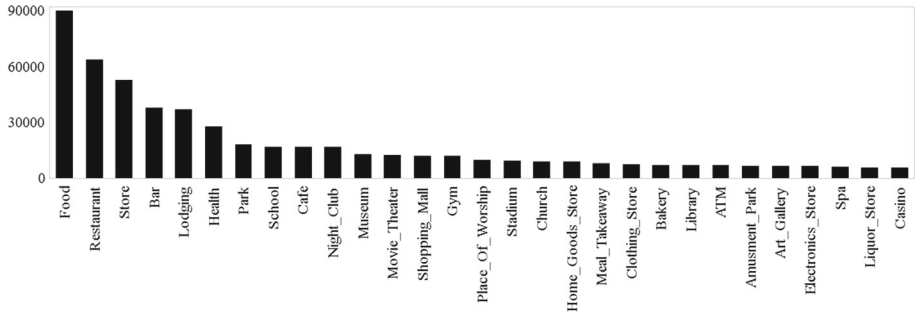

Fig. 8 Distribution for some of the top categories in training data set

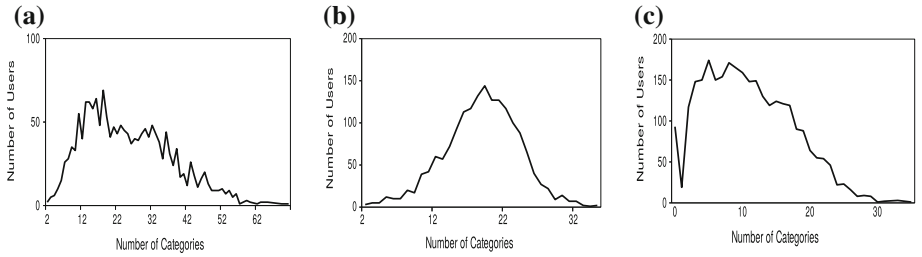

Fig. 9 Distribution of number of categories on data sets. **a** Training set. **b** Test set 1. **c** Test set 2

cafe, synagogue, church, florist, natural_feature, shopping_mall, grocery_or_supermarket, car_rental, convenience_store, locksmith, amusement_park, electronics_store, library, fire_station, gym, jewelry_store, mosque, campground, furniture_store, gas_station, meal_takeaway, clothing_store, train_station, bowling_alley, atm, parking, department_store, airport, movie_theater, electrician, real_estate_agency, food, pet_store, rv_park, doctor, zoo, shoe_store, laundry, recreation_center, discotheque.

The distribution for some of the top categories in training data set is shown in Fig. 8. You may note that the distribution of categories is very skewed.

3.2 Computing for additional constraints

We also understand the number of location categories that users visit on average. The higher the number, the higher the complexity of the prediction problem. Figure 9 shows the distributions for Training Set, Test Set 1, and Test Set 2. You may observe that the mode of the distributions corresponding to Training Set and Test Set 1 is 17 and that of Test Set 2 is 5.

Considering radius We have used GPA for finding out places near the target *location tweet*'s geo-coordinates. We can fetch approximately 60 places around the geo-coordinate with in specified radius in meters (for example: 50, 300, 1000, 2000). For each place L, we have set of categories L_c. Therefore, total number of categories around the geo-coordinate are union of L_{c_i}, where i is the i_{th} place around the geo-coordinate. We have done exhaustive experiments for predicting place of visit among these union of categories by setting different radii.

Considering time As mentioned in Eq. (4), we use $P(C_i/t)$, the probability of any person visiting place C_i at a given time t. To estimate, we need to find out the time of the day at which user visited a given place. The time stamps of tweets are according to standard time (GMT).

We convert this standard time into local time of that place from where tweet is generated. This involves finding the timezone of the place from where tweet is generated and then computing the local time of the tweet. We discretize the time into 1 h slots for computing the estimates. Let $n(C_i, t)$ be the number of times users visited the place C_i at time t, then:

$$P(C_i/t) = \frac{n(C_i, t)}{\sum_{\forall j} n(C_j, t)} \tag{5}$$

3.3 Evaluation method

Since we have the ground truth for test sets as well, we compute the accuracies of models as follows: (a) For every test instance we consider only top N categories returned by used model. (b) If ground truth is recovered from top N categories, then we considered that test instance as accurately classified. Therefore, if total number of test instances are M, and number of accurately classified test instances are K, then the prediction accuracy of model can be defined as follows.

$$Acc@N = \frac{K}{M}$$

We denote various models mentioned in Sect. 2 as follows:

(1) P(C/SystemU): Predictions using binary relevance method (BRM) on features formed using System U. We have used MULAN implementation [27] of BRM.
(2) P(C/Win): Predictions using Labeled LDA over the text in the historical window of tweets. We have used JGibbLabeledLDA [24] implementation to build the model. All the parameters of the algorithm are set to their default values except the number of most likely words for each topic. We set this parameter to 100.
(3) P(C/Win,t): Predictions using Eq. (4) in which $P(C_i/Win)$ is computed as in (2). Please refer to Sect. 3.2 for more details.

We compare our results with four baselines which are explained as follows. We want to mention that we are using these baselines, as there is no relevant work present in state of art to the best of our knowledge. These baselines are defined to highlight the non-obvious nature of the proposed approach.

– *Baseline Model 1* (Nearest): We considered categories in ascending order according to the distance from the target *location tweet*'s geo-coordinates. It is used to show the preference of user from query point according to distance.
– *Baseline Model 2* (Google Popularity): In this baseline, we considered the categories in the order returned by GPA. GPA returns places in the order of their popularity which is calculated using reviews by users on a given place.
– *Baseline Model 3* (P(C)): In this baseline, we have considered the popularity of a category in the training data. Let $n(C_i)$ be the number of times users visited the category C_i in the training set, Then,

$$P(C_i) = \frac{n(C_i)}{\sum_{\forall j} n(C_j)} \tag{6}$$

– *Baseline Model 4* (P(C/t)): We use estimates computed as mentioned in Eq. (5) to calculate the accuracies given only the time.

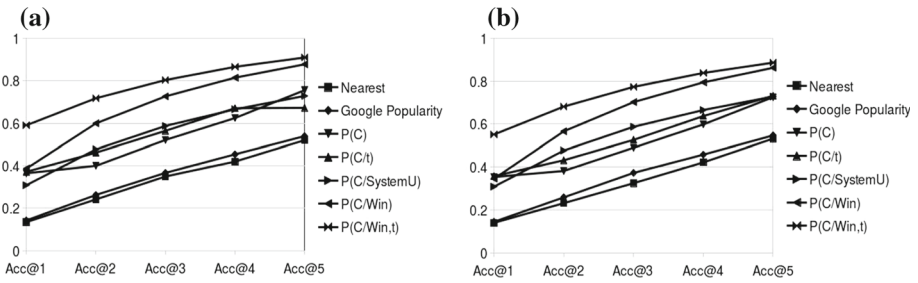

Fig. 10 Performance of $P(C/Win)$ on different window size. **a** Results on test set 1 (Radius: 300; Window size: 24 h). **b** Results on test set 2 (Radius: 300; Window size: 24 h)

3.4 Results

We conduct three sets of experiments. Section 3.4.1 presents the analysis of first set of experiments, where we compare the performances of various models in the proposed approach with those of the four baselines. In the second set of experiments performed in Sect. 3.4.2, we evaluate the impact of user history on performance of the proposed approach. In Sect. 3.4.3, third set of experiments evaluates the performance of the proposed approach with respect to various radii of the search space. The objective is to understand the behavior of the proposed approach under various circumstances.

3.4.1 Comparison with baselines

The experiments in this subsection are conducted using window length of 24 h, i.e., considering all tweets within 24 h before *location tweet*. This window size is chosen mainly because System U requires 100 words to profile a user. The accuracies of various algorithms are shown in Fig. 10, where Acc@N represents the accuracy using top N categories given by the algorithm.

The performance of our proposed approach is better that of all four baselines. The results show that models $P(C/Win)$ and $P(C/Win, t)$ have performed consistently better than $P(C$/SystemU) and four baselines across top 1 to top 5 predictions. These results indicate that the historical tweets by people are indicative of their immediate future activity. We have extracted the frequent words that people use in their tweet history before visiting different categories of locations. These words are shown in Table 3. The categories in the first column of the table are visited by the users, and the indicative words are those words which are extracted from their historical tweets just before visiting these categories. We can see that most of the words are indicative of their intent and present state of mind along with their activity performed after that.

Experiments also validate that, our proposed approach can be applied to wide variety of people because of its performance on Test Set 2. Hence, this could be the right solution to the cold start problem in a very general and simplified sense.

We observed that using only System U features for predicting categories is not useful, in contrary to the usefulness of System U features for people recommendations [4].

We also observe that using all words available on timelines are more predictive in nature in comparison to technique where some predefined set of words are used for deriving personality traits (example Linguist Inquiry and Word Count (LIWC), System U). The reason would be that using all words on timeline explores more latent factors that are not captured by predefined

Table 3 Word category relation captured by labeled LDA model

Category	Indicative keywords
Restaurant	Dinner, diner, coffee, food, lunch, donuts, dunkin, bakery, sushi, burger, casino, kitchen, night, orange, family, haven, reached, happy, fire, great, deli
Food	Love, time, like, good, today, photo, will, night, great, happy, tonight, thanks, work, going, best, birthday, well, next, ready, friends, drinking
Movie_theatre	Others, just, center, time, union, theater, love, night, posted, show, best, still, arts, movie, party, well, amazing, nice, ready, watch, performing
Health	Time, fitness, workout, gym, club, life, well, hill, great
Church	Church, toms, cancer, saint, massapequa, hall elementary, presbyterian, sagittarius, baptist, mary, fields, might, cathedral, citadel, christ, light, please, prayers, recovery
Park	Morning, walk, nice, work, better, river, green, summer, music, live
Shopping_mall	Shopping, mall, shop, next, best, grand, free, sale, nice

set of words. We have done our experiment with and without using stop words and found out that including stop words are giving little better results.

Also we can see that besides the most preferable categories by people, time plays bigger role in deciding the activity. This can inferred by comparing the results of $P(C)$ with $P(C/t)$. Therefore, while predicting location we used time as additional constraint with past tweets. Results have shown that by considering time, accuracy of model can be enhanced. This can be observed by comparing $P(C/Win)$ and $P(C/Win, t)$, where $P(C/Win, t)$ has given better results.

3.4.2 Performance with respect to various lengths of windows of historical tweets

In this experiment, we considered different window sizes that are 6, 12, 24, 36 h. Results are shown in Fig. 11, where WinN represents the size of window, N $\in$ {6 12 24 36}. We find that the optimum length of the window is 12 h in both Test Set 1 and Test Set 2 for better prediction. By observing the text of window carefully we found that when size of the window is greater than 12 h, sometimes tweets make less sense in predicting the place of visit. For example tweet like "I am feeling hungry" tweeted 24 h ago has nothing to do while predicting place of visit now. Contrary to this, some tweets like "Feeling bored, looking for some fun" tweeted 12 h ago could help in predicting place of visit now. On the other hand, while considering window size of 6 h, we found that sometimes there are very few tweets in window (i.e., 2 or 3), which made erroneous prediction because of very few words. By these results analysis we found that the more the recent information of user we have, the better the results will be in predicting place of visit for the user. The model used here is $P(C/Win)$.

3.4.3 Performance with respect to various radius of queries

In third set of experiments, we consider $P(C/Win)$ and $P(C/Win, t)$ for different radii—50, 300, 1000, 2000 (in meters). Figure 12 demonstrate results for different models. The overall performance trends are the same for Test Set 1 and Test Set 2. $P(C/Win, t)$ is performing better at radius 300 in comparison to all other spatial ranges considered in experiment except for the top 1 suggestion. Similarly prediction accuracy of $P(C/Win)$ is also highest except

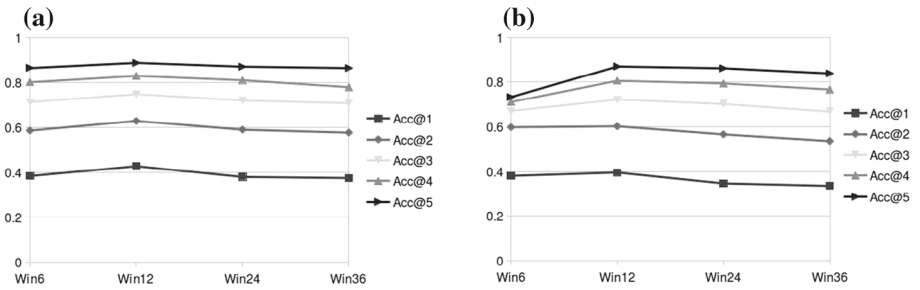

Fig. 11 Performance of $P(C/Win)$ on different window size. **a** Results on test set 1 using radius 300 m. **b** Results on test set 2 using radius 300 m

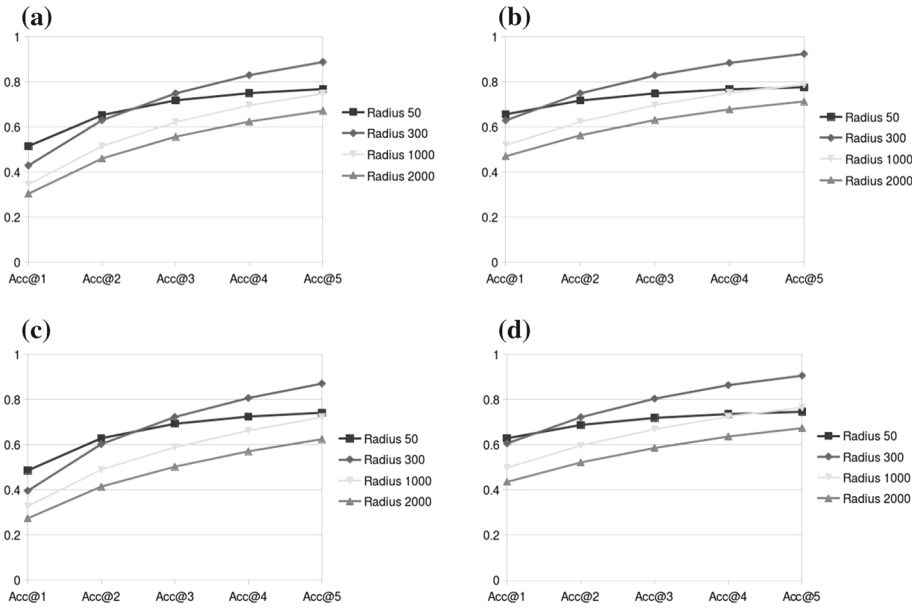

Fig. 12 Performance of models on various spatial ranges (i.e., Radius) by given models. **a** Performance of $P(C/Win)$ on test set 1. **b** Performance of $P(C/Win, t)$ on test set 1. **c** Performance of $P(C/Win)$ on test set 2. **d** Performance of $P(C/Win, t)$ on test set 2

for top 1 and 2 suggestions at the same radius i.e., 300 m. We analyzed that with the increase in spatial range more categories appear as candidate places of visit for user. Many times likelihood of distant places are greater than nearer places. Therefore while ranking according to likelihood for top N predictions, most of the distant places appeared. As we noticed that user generally preferred near by location (approx. within 300 m) from the search point (geo-coordinate) in data set. Hence results in erroneous predictions. For better understanding of above situation lets take an example. If user is more interested in watching movie since morning, then most of her tweets contain information relevant to movie only. However in evening, if she discovers that she has to spend large amount of time in traveling to movie theater because of distance/traffic, then she would prefer activity in her approach. Hence, tweeted less about that activity in comparison with movie due to less time. Now when we predict future places of visit for this user using her past 12-h tweets, we get higher

confidence of movie than preferred activity because of difference in activity relevant tweets. We empirically found out that the optimum scope of search should be optimal (300 m in this case) for better prediction. It should not be very small or very large.

4 Related work

We categorize the literature related to our work into three buckets. They are (a) Twitter data for prediction. (b) Predicting Locations Using LBSN. (c) Deriving personality traits from microblog text for prediction. We review these efforts in the following subsections and contrast with the proposed work wherever it is appropriate.

4.1 Twitter data for predictions

Twitter data has been widely explored for predictions. Yuan et al. [40] proposed a probabilistic model to discover individual users' mobility behavior from spatial, temporal and activity aspects using twitter data. Lichman et al. [23] model human location data with mixture of kernel densities and predict a spatial distribution for an individual. In comparison to our work, the model proposed by Lichman and Smyth [23] works for only those users which are seen by model and give the likelihood probability of user at given coordinate. Sadilek et al. [33] proposed a system nEmesis, which is used to identify the restaurants not to be visited. nEmesis ranked restaurant as negative or positive by tracking the activities of user after that restaurant visit. Abel et al. [1] proposed a frame work in which user timelines from twitter are used to model the user profile. Results shows that by considering temporal dynamics and exploiting tweet-news relationship better recommendation systems can be made. Similarly [19,28] also proposed the system for new recommendation by using twitter. Buza et al. [10] have shown that their proposed approach may be useful for prediction tasks in the financial domain. For this they collected only financial tweets and tried to predict the yield of particular stocks based on those tweets. Similar to our work, Ritterman et al. [32] have shown that Twitter encodes the belief of people about some concrete statement about the world. They used these beliefs with prediction market to predict a swine flu pandemic.

In contrast to our approach where we predict POI such as shop, church, restaurant based on intent of user by using past vocabulary, Han et al. [17] predict user location (i.e., country or region) by identifying location indicative words (i.e., frequent word used at specific location). Lee et al. [21] used Foursquare to predict the location of tweets. They build probabilistic models using unstructured text coupled with semantic locations. Ramasamy et al. [31] tried to infer the user interest from tweet times. Simple heuristic behind this work is that user often tweet at larger rate during the happening of event in interest, in respect to non-event time. Which shows their interest in that particular event. Similarly Budak et al. [9] address the problem of inferring user interests from Twitter based on their utterances and interaction in the social networks. Novel probabilistic generative model is proposed to based on user utterances that encapsulate both user and network information. Bhatacharya et al. [6] infer users' topic of interest by the followee list of twitter users'. They observe that user generally follow experts on various topics of her interest. By deducing the followee expertise area, user interest can be inferred.

Bollen [8] study the correlation of public mood derived from large scale collection of tweets to stock market index. Results has shown that prediction accuracy can be improved significantly by the inclusion of public mood dimension. Asur et al. [3] predicts box-office revenues of movies in advance of their release by using tweets using linear regression

model which out performs Hollywood stock exchange predictions. In contrast, Gayo-Avello et al. [13] contradicts the predictive power of tweets. In their study, they contradict the previously published results that claims that predicting electoral outcomes from social media data is feasible.

Blogs used for prediction Yin et al. [39] proposed a model where each region is characterized by a topic distribution, and represented by a bi-variant Gaussian distribution over coordinates. Modeling geographical topics includes location-driven model, text-driven model, and Latent Geographical Topic Analysis that combines both location and text information. Both [18,38] mine the relationship between location and words usage at that location using probabilistic graphical models.

4.2 Predicting locations using LBSN

Mathew et al. [26] predict future location of an individual by capturing the sequential relations between places visited in given time period by using Hidden Markov Models. In this approach locations are clustered according to their characteristics and latter used for training an HMM for each cluster. Bao et al. [5] proposed location-based and preference-aware recommendation system using sparse geo-spatial networking data from Foursquare without considering temporal information. This approach is similar to our proposed approach as they have considered categories for recommendation except that [5] could not address the cold start problem in case of new users. Yuan et al. [42] proposed a graph-based point-of-interest recommendation system using temporal and geographical influences. In given approach, Yuan et al. [42] have used category of place rather than the name of place considering temporal factor, though data is used from Foursquare. For point-of-interest recommendation, algorithm has been proposed based on following observations (a) users tend to visit nearby places (b) users' visit different places at different time slots. But contrary to our approach this system recommend only if the user history is available but not for new user (cold start problem). Yuan et al. [41] proposed a system using collaborative filtering that exploits both temporal and spatial information specific to point of interest for recommendations for each individual. Gao et al. [12] used matrix factorization technique for exploring temporal effects on location-based social networks for recommendations which again results in cold start problem for new place and people.

4.3 Deriving personality traits from microblog text for prediction

Lexicon used by person is a representation of his personality which is confirmed by research work done in last decade [20,29,35]. Common approach of analyzing the personality of person is to count the words falling in particular category given by Linguist Inquiry and Word Count (LIWC) [29].

Chen et al. [11] used social media word for understanding individuals' personal values. They have shown that personal values can influence word usage and also words usage contains predictive information for person's values which can be used for other meaningful prediction tasks. Golder et al. [15] used million of tweets for finding out the diurnal and seasonal mood rhythms at individual level. In their study they have shown that by lexical analysis we can predict the state of mind of individual such happy, sleepy. Study has shown that age and gender of person can be inferred by analysis the usage of words in their utterances over blog, emails, social networking sites etc. [2,14]. Mahmud et al. [25] proposed a model that can identify the strangers over twitter who are eligible and willing to provide the required information. Lee et al. [22] addressed the problem of identifying the stranger on twitter for propagating the

information by re-tweeting the desired tweet. Badenes et al. [4] also used tweets for deriving personality traits by using System U and then used them for people recommendations.

All these approaches has been used extensively in analyzing personality traits, but it has also shortcomings of predefined word category correlation.

Open-vocabulary approach Schwartz et al. [34] used rather different approach for using vocabulary know as open-vocabulary technique (using all set of words available on social media) in comparison with closed-vocabulary technique, Linguistic Inquiry and Word Count (LIWC) [29], where some predefined set of words are used for deriving the personality traits. In their study they have shown that by using open-vocabulary approach they got higher state-of-art accuracy in predicting gender in comparison to LIWC by exploring latent factors that are not captured by closed-vocabulary approach. Motivated by this approach we have used all words available on timelines for modeling the users' behavior.

5 Conclusion

This paper presents prediction of users' location using their twitter handle. We identified tweets within user timelines that contain location information that the user would have visited. Thus, we have created a large amount of training data. We extracted two types of features from the past tweets viz., System U profiles and actual text in a window of historical tweets. We found out that, using System U profiles only are not very encouraging for predicting locations. Text from the tweets preceding a location tweet are concatenated, labeled with the category of the location and this data used to train Labeled LDA. We found that Labeled LDA captures the relationship between the words used by user and location visited by him in future. This also validate our hypothesis that vocabulary can be used to capture the intent and present state of mind which majorly decides the activity done by user in future. One of the major contributions of our work is in finding the sequence of places people have visited using their twitter timeline. This has helped us gain many insights into behavior of various prediction models and the associated parameters. We compared the performance of the proposed approach with four baselines. The models in our approach performed better than all these four baselines. The highest accuracy of one of the proposed model is over 90 % for the top five suggestions. We found that the behavior of the proposed approach is very similar to both seen and unseen users. This indicates a solution to cold start problem in very unique way, both for new users or location which are unseen by the model.

We conclude this paper by explaining a few technical challenges relating to the proposed work. We have tested our approach on a random sample of people. Can this approach work for all those that we want to predict locations for? In order words, how do we identify class of people that a single model can be applied to?

Some of the immediate extensions to the present work that we are looking at are as follows. People may start tweeting for longer duration before visiting a place of higher significance in their lives. For example, people may tweet about attending a soccer match for over a month before the match, whereas they may tweet only a day before visiting a good restaurant. In this paper, we consider a fixed window of past tweets for all the categories. It will be interesting to design classifiers that can consider category-adaptive size of window of tweets. Also, we assume in the current work the next location of visit is independent of the current activity/location. However, the current place of a person could influence where she visits in the next time interval. For example, the likelihood of a person going on a roller-coaster ride immediately after a meal is very small. It will be interesting to infer the

dependence of past locations on the future locations from the training data and use it in generating recommendations. Moreover, in this work we predict the place of visit assuming the knowledge of user's current location at given time but it would be more interesting to first estimate the user's location after a given time and then predict the place of visit around that estimated location. Further to show the predictive power of tweet for predicting place of visit of user, we use different sizes of widow, but for the real time system the appropriate size of the window should be determined using the training data only which is also include in our future work. Also, scalability of the system is also to be explored on large level like considering users of different countries or continents at a time. Because sometimes people at different location use different type of dialects for the same purpose and vice versa which may effect the performance of the system.

References

1. Abel F, Gao Q, Houben G-J, Tao K (2013) Twitter-based user modeling for news recommendations. In: Rossi F (ed) IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3–9, 2013. IJCAI/AAAI. http://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6683
2. Argamon S, Koppel M, Pennebaker JW, Schler J (2007) Mining the blogosphere: age, gender and the varieties of self-expression. First Monday **12**, 9. http://dblp.uni-trier.de/db/journals/firstmonday/firstmonday12.html#ArgamonKPS07
3. Asur S, Huberman BA (2010) Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology—Volume 01 (WI-IAT '10). IEEE Computer Society, Washington, DC, USA, pp 492–499. doi:10.1109/WI-IAT.2010.63
4. Badenes H, Bengualid MN, Chen J, Gou L, Haber E, Mahmud J, Nichols JW, Pal A, Schoudt J, Smith BA, Xuan Y, Yang H, Zhou MX (2014) System U: automatically deriving personality traits from social media for people recommendation. In: Proceedings of the 8th ACM conference on recommender systems (RecSys '14). ACM, New York, NY, USA, pp 373–374. doi:10.1145/2645710.2645719
5. Bao J, Zheng Y, Mokbel MF (2012) Location-based and preference-aware recommendation using sparse geo-social networking data. In: Proceedings of the 20th International conference on advances in geographic information systems (SIGSPATIAL '12). ACM, New York, NY, USA, pp 199–208. doi:10.1145/2424321.2424348
6. Bhattacharya P, Zafar MB, Ganguly N, Ghosh S, Gummadi KP (2014) Inferring user interests in the Twitter social network. In: Kobsa A, Zhou MX, Ester M, Koren Y (eds) Eighth ACM conference on recommender systems, RecSys '14, Foster City, Silicon Valley, CA, USA—October 06–10, 2014, ACM, 357–360. doi:10.1145/2645710.2645765
7. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022. http://dl.acm.org/citation.cfm?id=944919.944937
8. Bollen J, Mao H, Zeng X-J (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8. doi:10.1016/j.jocs.2010.12.007
9. Budak C, Kannan A, Agrawal R, Pedersen J (2014) Inferring user interests from microblogs. Technical Report MSR-TR-2014-68. http://research.microsoft.com/apps/pubs/default.aspx?id=217311
10. Buza K, Nanopoulos A, Nagy G (2015) Nearest neighbor regression in the presence of bad hubs. Knowl Based Syst 86:250–260. doi:10.1016/j.knosys.2015.06.010
11. Chen J, Hsieh G, Mahmud J, Nichols J (2014) Understanding individuals' personal values from social media word use. In: Fussell SR, Lutters WG, Morris MR, Reddy M (eds) Computer supported cooperative work, CSCW '14, Baltimore, MD, USA, February 15–19, 2014, ACM, pp 405–414. doi:10.1145/2531602.2531608
12. Gao H, Tang J, Hu X, Liu H (2013) Exploring temporal effects for location recommendation on location-based social networks. In: Yang Q, King I, Li Q, Pu P, Karypis G (eds) Seventh ACM conference on recommender systems, RecSys '13, Hong Kong, China, October 12–16, 2013, ACM, pp 93–100. doi:10.1145/2507157.2507182
13. Gayo-Avello D, Metaxas PT, Mustafaraj E (2011) Limits of electoral predictions using twitter. In: Adamic LA, Baeza-Yates RA, Counts S (eds) ICWSM, The AAAI Press. http://dblp.uni-trier.de/db/conf/icwsm/icwsm2011.html#Gayo-AvelloMM11

14. Gilbert E (2012) Phrases that signal workplace hierarchy. In: Poltrock SE, Simone C, Grudin J, Mark G, Riedl J (eds) CSCW, ACM, 1037–1046. http://dblp.uni-trier.de/db/conf/cscw/cscw2012c.html#Gilbert12
15. Golder SA, Macy MW (2011) Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science 333(6051):1878–1881. doi:10.1126/science.1202775
16. GoogleAPI (2015) Google Places API. https://developers.google.com/places/documentation
17. Han B, Cook P, Baldwin T (2014) Text-based twitter user geolocation prediction. J Artif Intell Res 49:451–500. doi:10.1613/jair.4200
18. Hao Q, Cai R, Wang C, Xiao R, Yang J-M, Pang Y, Zhang L (2010) Equip tourists with knowledge mined from travelogues. In: Rappa M, Jones P, Freire J, Chakrabarti S (eds) In: Proceedings of the 19th international conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26–30, 2010, ACM, pp 401–410. doi:10.1145/1772690.1772732
19. Jonnalagedda N, Gauch S (2013) Personalized news recommendation using twitter. In: IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT), vol 3, pp 21–25. doi:10.1109/WI-IAT.2013.144
20. Kramer ADI, Chung CK (2011) Dimensions of self-expression in facebook status updates. In: Adamic LA, Baeza-Yates RA, Counts S (eds) ICWSM, The AAAI Press. http://dblp.uni-trier.de/db/conf/icwsm/icwsm2011.html#KramerC11
21. Lee K, Ganti RK, Srivatsa M, Liu L (2014a) When twitter meets foursquare: tweet location prediction using foursquare. In: Proceedings of the 11th international conference on mobile and ubiquitous systems: computing, networking and services (MOBIQUITOUS '14). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, pp 198–207. doi:10.4108/icst.mobiquitous.2014.258092
22. Lee K, Mahmud J, Chen J, Zhou MX, Nichols J (2014b) Who will retweet this? automatically identifying and engaging strangers on twitter to spread information. http://arxiv.org/abs/1405.3750
23. Lichman M, Smyth P (2014) Modeling human location data with mixtures of kernel densities. In: Macskassy SA, Perlich C, Leskovec J, Wang W, Ghani R (eds) The 20th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '14, New York, NY, USA, August 24–27, 2014, ACM, pp 35–44. doi:10.1145/2623330.2623681
24. Labeled LDA (2015) Labeled LDA in Java. (2015). https://github.com/myleott/JGibbLabeledLDA
25. Mahmud J, Zhou MX, Megiddo N, Nichols J, Drews C (2013) Recommending targeted strangers from whom to solicit information on social media. In: Kim J, Nichols J, Szekely PA (eds) 18th International conference on intelligent user interfaces, IUI '13, Santa Monica, CA, USA, March 19–22, 2013, ACM, pp 37–48. doi:10.1145/2449396.2449403
26. Mathew W, Raposo R, Martins B (2012) Predicting future locations with hidden Markov models. In: Dey AK, Chu H-H, Hayes GR (eds) The 2012 ACM conference on ubiquitous computing, Ubicomp '12, Pittsburgh, PA, USA, September 5–8, 2012, ACM, 911–918. doi:10.1145/2370216.2370421
27. MLib (2015) MULAN java library. (2015). http://mulan.sourceforge.net
28. De Francisci Morales G, Gionis A, Lucchese C (2012) From chatter to headlines: harnessing the real-time web for personalized news recommendation. In: Adar E, Teevan J, Agichtein E, Maarek Y (eds) Proceedings of the fifth international conference on web search and web data mining, WSDM 2012, Seattle, WA, USA, February 8–12, 2012, ACM, pp 153–162. doi:10.1145/2124295.2124315
29. Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ (2007) The development and psychometric properties of LIWC2007. Austin, TX, LIWC. Net (2007)
30. Ramage D, Hall David LW, Nallapati R, Manning CD (2009) Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on empirical methods in natural language processing, EMNLP 2009, 6–7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL, pp 248–256. http://www.aclweb.org/anthology/D09-1026
31. Ramasamy D, Venkateswaran S, Madhow U (2013) Inferring user interests from tweet times. In: Muthukrishnan SM, Abbadi AEl, Krishnamurthy B (eds) Conference on online social networks, COSN'13, Boston, MA, USA, October 7–8, 2013, ACM, pp 235–240. doi:10.1145/2512938.2512960
32. Ritterman J, Osborne M, Klein E (2009) Using prediction markets and twitter to predict a swine flu pandemic. In: Proceedings of the 1st international workshop on mining social media. http://www.socialgamingplatform.com/msm09/proceedings/paper2.pdf
33. Sadilek A, Brennan SP, Kautz HA, Silenzio V (2013) nEmesis: which restaurants should you avoid today? In: Hartman B, Horvitz E (eds) HCOMP, AAAI. http://dblp.uni-trier.de/db/conf/hcomp/hcomp2013.html#SadilekBKS13
34. Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, Shah A, Kosinski M, Stillwell D, Seligman ME (2013) Ungar LH (2013) Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS One 8:9. doi:10.1371/journal.pone.0073791

35. Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 29(1):24–54. doi:10.1177/0261927X09351676
36. Tsoumakas G, Katakis I, Vlahavas I (2010) Mining multi-label data. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook, Springer US, pp 667–685. doi:10.1007/978-0-387-09823-4_34
37. TwAPI (2015) Twitter streaming api. https://dev.twitter.com/docs/using-search
38. Wang C, Wang J, Xie X, Ma W-Y (2007) Mining geographic knowledge using location aware topic model. In: Proceedings of the 4th ACM Workshop on Geographical Information Retrieval. GIR '07. ACM, NY, USA, pp 65–70. doi:10.1145/1316948.1316967
39. Yin Z, Cao L, Han J, Zhai C, Huang TS (2011) Geographical topic discovery and comparison. In: WWW. pp 247–256
40. Yuan Q, Cong G, Ma Z, Sun A, Magnenat-Thalmann N (2013a) Who, where, when and what: discover spatio-temporal topics for twitter users. In: Dhillon IS, Koren Y, Ghani R, Senator TE, Bradley P, Parekh R, He J, Grossman RL, Uthurusamy R (eds) The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11–14, 2013, ACM, pp 605–613. doi:10.1145/2487575.2487576
41. Yuan Q, Cong G, Ma Z, Sun A, Thalmann NM (2013b) Time-aware Point-of-interest recommendation. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval (SIGIR '13). ACM, New York, NY, USA, pp 363–372. doi:10.1145/2484028.2484030
42. Yuan Q, Cong G, Sun A (2014) Graph-based Point-of-interest recommendation with geographical and temporal influences. In: Li J, Wang XS, Garofalakis MN, Soboroff I, Suel T, Wang M (eds) Proceedings of the 23rd ACM international conference on conference on information and knowledge management, CIKM 2014, Shanghai, China, November 3–7, 2014, ACM, pp 659–668. doi:10.1145/2661829.2661983

Arun Chauhan is a Research Scholar at department of Computer Science and Engineering, Indian Institute of Technology, Roorkee. His Ph.D. thesis topic is Recommendation Systems Using Microblogs. He received his M.Tech. degree from Indian Institute of Technology, Kharagpur in 2010. He has teaching experience of more than 5 years. He authored and co-authored few papers in field of Speech Processing also during his M.Tech. at IIT Kharagpur. His research interest are in Data Mining, Social Media Analysis, Big Data Analysis and Machine Learning.

Krishna Kummamuru is a senior researcher at IBM India Research Lab working on solutions for financial industry. In this past, he worked as a software architect developing IBM Watson based cognitive solutions and contributed to WEA product. He led Collaboratory for Service Science at ISB, Hyderabad. He had actively contributed to several Research Accomplishments working on projects relating to machine learning. He received B.Sc degree from Nagarjuna University in 1986. He received the M.E degree and the Ph.D. in Electrical Engineering from IISc, Bangalore, India in 1993 and 1999 respectively. He received Alfred Hay Medal for the best graduating student in EE in 1993 from IISc. He has 13 patents granted by USPTO (filed a total of about 30 patents). He has over 35 publications in refereed conferences and journals (a citation count of over 1600). His technical interests include Machine learning, Text mining, Recommendation systems and User interfaces.

Durga Toshniwal is presently working as Associate Professor at the Department of Computer Sceience and Engineering, Indian Institute of Technology Roorkee, India. She completed her Bachelor in Engineering and subsequently earned her Master of Technology from National Institute of Technology, Kurukshetra and Doctor of Philosophy from Indian Institute of Technology Roorkee, India. Some of her areas of research interests include Data Mining, Social Media Analysis, Big Data Analysis and Machine Learning. Dr. Durga has published her research work in several international journals and conferences. Dr Durga has received various awards and honours. Some recent ones are—IBM Faculty Award, Award from UNESCO Chair in Data Privacy, and the very prestigious IBM Shared University Research Award 2009 for her research projects.