Using Twitter for Next-Place Prediction, with an Application to Crime Prediction

Mingjun Wang and Matthew S. Gerber
Department of Systems and Information Engineering
University of Virginia
Charlottesville, Virginia 22903
Email: {mw4cc, msg8u}@virginia.edu

***Abstract*—This research focuses on two problems. First, we investigate the prediction of social media users' spatial trajectories. Recent work on this task has focused on the use of cellular network traces and location-based social network services such as Foursquare, all of which emit structured geospatial information (e.g., cellular tower identifiers, GPS coordinates, and venue identifiers). Less attention has been paid to the rich textual content that users often publish in tandem with the structured information. We investigate methods of integrating textual content into existing next-place prediction models, and we demonstrate a significant improvement in next-place prediction compared to several baselines derived from published research. Second, we examine the correlation between these next-place predictions and the occurrence of crimes in a major United States city, with the goal of aiding future research into automatic crime prediction.**

I. Introduction

Crimes are more likely to happen at the space-time confluence of attackers, victims, and absence of protective elements [23], [28]. Thus, being able to predict individuals' movement patterns could be a useful aspect of effective crime prediction and policing. This paper presents research on automatically predicting users' spatial trajectories, a problem known as next-place prediction [1]. Services such as Foursquare and Facebook Places allow users to "check in" at venues and broadcast this information to their social network. Most recent work on next-place prediction has used these check-ins to predict users' next-place trajectories. This work, which will be reviewed in more detail in Section II, has largely ignored the rich textual content of social media posts, which we hypothesize can substantially improve the accuracy of next-place prediction models. Specifically, our first hypothesis is as follows:

- H1: An individual's future venue trajectory correlates with his or her historical tweets.

Twitter has strong potential in predicting and describing election results [29], natural disasters [30] and crime [15], [32]. Thus, we propose to investigate the correlation between our text-enriched next-place predictions and the occurrence of crimes. Specifically, our hypothesis is as follows:

- H2: Crime rates correlate with the density of users' movement trajectories in the same area.

The primary challenge in using Twitter for next-place prediction and crime prediction is that the textual content of tweets does not typically bear any overt connection with geospatial locations. Occasionally, a user will mention a specific address or business, or the user will attach a Foursquare check-in to their tweet. More often, indicators of future movement are left implicit. Consider the following tweets, which demonstrate these cases:

- Mention of a specific address: Thanks to the "lady" at 3737 North Western Avenue, I'll never order Pete's Pizza again. And @GrubHub, done with you.
- Foursquare check-in: Check out ESPNChicago - #StateStreetStudio (190 N State St, at Lake St, Chicago) on @foursquare: http://t.co/PnFFbkTQ3m
- Implicit trajectory hint: @joshua_ocampoo: I'm hungry

The primary contribution of this research is to address posts such as the third one above (implicit). In this case, we are uncertain about the user's future spatial trajectory, but the textual content presents information that might imply movement of the user from his or her current location to a local dining establishment.

We present two models for text-enriched next-place prediction. The first predicts the type of venue (e.g., restaurant or transport hub) the user will visit next. The second predicts how far the user will be from each type of venue (e.g., 100m from a restaurant and 3500m from a transport hub). We then test the correlation between predicted concentrations of users at these venue types and the occurrence of future crimes at such venues. Thus, our contribution is twofold:

- We develop and test a text-enriched model for next-place prediction based on social media posts.
- We formally test the correlation between next-place concentrations and the occurrence of actual future crimes in a large United States city.

This paper is structured as follows: in Section II, we review research on next-place and crime prediction. In Section III, we describe our data sources and preparation steps. In Sections IV and V, we describe our models and results for next-place prediction. In Section VI, we examine the correlation between next-place trajectories and crime, and we move toward a better crime prediction model in Section VII. In Section VIII, we conclude with a summary of our research and ideas for future work.

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.138

IEEE computer society

II. Related Work

A. Next-Place Prediction

In general, there are two types of work in this area. First is the prediction of an individual's home location, motivated by the fact that few users post their locations in social media [9]. The current work in home prediction uses content from social media check-ins [9], [18], [2]. Second is the prediction of an individual's location at any time, which is the focus of our work. Most work utilizing mobile network data such as location information from GPS sensors or WiFi focuses on movement trajectories. There are a variety of applications for next-place prediction, including mobile advertising [3] and disaster relief [17].

Researchers have investigated many explanatory variables for next-place prediction, including cell phone data usage [24], [21], [10], visiting frequency and contextual information from smart phone sensors [11]. Also, with the rapid growth of location-based social networks (LBSN), researchers have used check-in patterns to predict the next check-in. In [25], researchers further investigated the problem by using a set of features describing users' movement patterns. Moreover, researchers have measured and compared the similarity between different users in social media for next-place prediction by collaborative filtering [20].

Researchers have proposed a variety of algorithms for next-place prediction, but most focus on classification [21] and Markov-based models [10]. In order to more carefully consider movement patterns, [12] extend the traditional Markov model to Mobility Markov Chain which demonstrates improved predictive performance. Researchers have investigated the use of social ties in social media to predict check-in patterns [13]. Researchers have also developed location-based recommendation systems using venue review and check-in histories [19]. Lastly, researchers have incorporated time into spatial prediction models [27], [14].

B. Crime Prediction

The criminological theory of routine activities suggests that crime is likely to occur at the space-time confluence of offenders, targets, and the absence of capable guardians [28]. Traditional hot-spot maps [6] produce retrospective visualizations of crimes, which can be predictive of future crime in cases where the occurrence of crime is stationary in space. Other research [15] integrates layers of geospatial information with historical crime records to improve these predictions. The present research provides preliminary evidence that these models could be further improved by including predicted concentrations of users at various venue types.

III. Data Preparation

We choose to use Twitter content for next-place prediction because Twitter users often provide clues about their daily activities using this platform. We prepared the data in three steps: tweet collection, tweet preprocessing, and matching tweets with Foursquare venues.

TABLE I. Distance between each geotagged tweet and their nearest venue (m)

Average	Standard Deviation	Max	Min
58.43	60.18	2768.47	0

A. Tweets and Foursquare Venues Collection

We extracted geotagged tweets from Twitter and obtained typed venue locations from Foursquare. We collected geotagged tweets with textual contents, user ID and geographical coordinates of longitude and latitude through Twitter's streaming API. All geotagged tweets are taken within the city boundary of Chicago, Illinois, USA in January 2014. We retained users who posted at least 20 tweets in this month. Thus, there are 1,233,076 tweets from 9,567 users in our data set. The venues in Chicago are extracted from check-in histories on Foursquare, which include the following ten categories: Travel & Transport, Food, Residence, Outdoor & Recreation, Professional & Other Places, Arts & Entertainment, Nightlife Spot, College & University, Shop Services and Event. In total, there are 224,124 venues in Chicago.

B. Tweet Preprocessing

We first apply part of speech (POS) tagging using the TweetNLP tool configured with 25 coarse POS tags [16]. We apply Tweet NLP to tokenize the textual content and tag the tokens. Since tweets are informal and short, we filter the text contents by removing the following parts-of-speech: determiner, postposition, coordinating conjunction, predeterminters, punctuation and numeral. Then we removed stop words according to the list provided by [5]. We also tagged @ symbols for replies and mentions in tweets, which identify social relationships within the data.

C. Matching Tweets with Foursquare Venues

To fill the gap between tweets and the physical environment for the next-place prediction problem, we propose two ways to anchor tweets to the physical environment:

- Nearest venue type (categorical)
- Minimum distance to each venue type (continuous)

We matched each observed geotagged tweet with the type of its nearest venue from Foursquare and calculated the distance between the tweet and venue. These values constitute the responses for our two next-place prediction problems (nearest venue type and distance to all venue types, respectively).

IV. Next-Place Prediction Problem

In this section, we present our approaches for incorporating textual content into next-place prediction models. When an individual posts a tweet, we define the next-place prediction problem as the task of predicting the location where this individual will post his or her next tweet. We formulate the next-place prediction problem in two ways: predicting the nearest venue type and predicting distances to each type of venue.

A. Text-Enriched Classification Model

Our text-enriched classification model predicts each user's nearest next venue type. The model has two parts. First is a binary classification model to determine whether an individual will maintain the current nearest venue type or move such that a different venue type becomes the nearest. Second, for users predicted to transition to a new venue type, we build a multivariate classification model to predict the venue type that will become nearest. The general forms of these models are:

$$Step1 : P(c_{n+1} = c_n|\chi) = F(f_1, f_2, ..., f_n),$$
$$Step2 : P(c_{n+1} = v|\chi) = F(f_1, f_2, ..., f_n)$$

where $\chi = c_1, c_2, ..., c_n$ is the user's visiting history and c_{n+1} is the next venue type. For each c_i in χ, c_i is a venue type, rather than a particular venue instance. Predictor variables $f_1, f_2, ..., \ f_n$ are the features extracted from historical venue trajectories and tweets associated with the user. We model each of the above classification problems with linear support vector machines implemented by LibLinear [7]. We have studied two classification models, which are the Text-Enriched Model and the Text-Enriched with @-link Model.

1) Text-Enriched Model: We extract features from the textual content of a user's tweets and the locations of these tweets to build the Text-Enriched Model.

- Hypothesis: Individuals' historical textual content correlates with his or her future venue trajectory.

Under this hypothesis, f_1 is the current venue type and $f_2, f_3, ..., f_n$ are TF-IDF features from textual content of the user's recent tweets so that the general forms of the models become

$$Step1 : P(c_{n+1} = c_n|\chi) = F(c_n, tfidf(t_n)),$$
$$Step2 : P(c_{n+1} = v|\chi) = F(c_n, tfidf(t_n)),$$

where χ is defined as the historical visiting history, c_{n+1} is the next venue type, t_n is the tweet posted from the current venue, and $tfidf(t_n)$ is the set of features extracted from the current tweet's textual content. Features extracted from textual contents are represented with Term Frequency Inverse Document Frequency (TF-IDF) as a vector space model, where the IDF component is calculated from all historical tweets.

2) Text-Enriched with @-link Model: The Text-Enriched with @-link Model extends the Text-Enriched Model with features extracted from tweets that mention the current user. This model starts with the feature set described above for the Text-Enriched Model, and it adds to this set features that capture the following hypothesis:

- Hypothesis: The current user's next location will correlate with the location of other users who have recently mentioned the current user.

Intuitively, users might travel to locations where their friends are located. We use the @-link in tweets to capture a rough notion of friendship. For each tweet, there are two groups of features. The first is the set used by the Text-Enriched Model, which includes the current venue type and the TF-IDF features. Features from the second group are extracted from other tweets that mention the current user: $f_{n+1}, ...f_{2n}$ are the venue type and TF-IDF features extracted from recent tweets that mention the current user. From all the venue types within tweets mentioning the current user, we use the venue type that is most recent as the value of f_{n+1}. The general form of this model as follows:

$$Step1 : P(c_{n+1} = c_n|\chi) = F(c_n, tfidf(t_n), m(u)),$$
$$Step2 : P(c_{n+1} = v|\chi) = F(c_n, tfidf(t_n), m(u)),$$

where m_u is set of features extracted from recent tweets that mention the current user.

B. Text-Retrieval Model

Intuitively, tweets posted near the same venue type should contain similar textual content. Thus, it should be possible to retrieve tweets corresponding to a venue type based on textual content. We build a collection of documents, one for each venue instance in the city. If we consider a user's current tweet as a query against this document collection, then the top-ranked document with respect to this query might be the user's targeted next place. For example, if the user mentions food in his or her tweet, this would presumably match tweets from food establishments, leading to a prediction of Food as this user's next venue type.

We implemented the above ideas with a simple vector space model of individual tweets (queries) and documents derived from tweets posted at particular venues. We used the BM25 [26] query model, which is defined as follows:

$$score(q, D) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_i + 1)}{f(q_i, D) + k_i \cdot (1 - b + b \cdot \frac{|D|}{avgdl}}$$

where q is the textual content of a tweet, and D is the collection of historical tweets in one venue type. We will get a score for the tweet and the document for each specified venue. For any words q_i in q, we get $f(q_i, D)$ as term frequency in the document and $IDF(q_i)$ is inverse document frequency for the query term. $\mid D \mid$ is the length of document D, $avggl$ is the average document length in all venues. k_1 and b are free parameters. Without further optimization, we set $k_1 = 1.5$ and $b = 0.75$ [22]. The venue type with the highest score will be the output of the Text-Retrieval Model.

C. Text-Enriched Regression Model

The features and formulation of the Text-Enriched Regression Model are similar to those of the Text-Enriched Classification Model, except that the response is vector of continuous distances, one for each venue type. The general form of the regression model is,

$$y = F(c_n, tfidf(t_n), m(u))\beta + \epsilon,$$

where,

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{10} \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{10} \end{pmatrix},$$

$$F(c_n, tfidf(t_n), m(u)) = \begin{pmatrix} F_1^T \\ F_2^T \\ \vdots \\ F_{10}^T \end{pmatrix}$$

The F_i^T are feature vectors capturing the user's nearest distance to each venue type and features from textual content. We build two regression models. The first is a main effects model that uses all of these features, and the second is a model that contains interaction terms for the venue distances.

D. Baseline Models

To evaluate the classification-based approaches for next-venue prediction, we built baseline models chosen from state-of-the-art published research.

1) Most Frequent Check-in Model: [8] showed that check-in frequency is the strongest predictor of users' next locations. We use this as our first baseline model. Considering each user's check-in history, the model generates a probabilistic distribution on venue types based on the visiting history. For the next location *v*, the probability is defined as,

$$P(c_{n+1} = v|\chi) = \frac{\#\ check-ins\ to\ v}{\#\ total\ check-ins}$$

2) Markov Model: Gambs et al. propose to address the problem by developing a Markov model to incorporate the k previously visited locations [12]. The Markov Model with order 2 performed the best in their experiments. In our experiments, we take both order-1 and order-2 as baseline models and the Most Frequent Check-in Model is the Order-0 Markov Model.

3) Classification Model with Historical Visiting Information: In [4], multiple algorithms were compared in the prediction of next-place with mobility data. The authors showed that the historical visiting frequency to each venue is useful for next-place prediction. Thus, we use a support vector machine to build a baseline classification model with features quantifying the historical visiting frequency to each venue type. The comparison between this classification model and our text-enriched model shows whether these features are useful.

Since the regression approach to distance prediction is new, a previously developed baseline does not exist. As a baseline for the regression models, we used the average distance from each venue type as the baseline prediction. We computed the average distance in the following way: We first calculate the distance to each venue type for all historical tweets. Then we get the average distance from the distances in these tweets.

V. Next-Place Prediction Results and Discussion

We measure the performance of our models as follows. For the prediction of nearest venue type, we define the prediction accuracy to be the ratio of the number of correct predictions and the total number of predictions:

$$PredictionAccuracy = \frac{Number\ of\ Correct\ Predcitions}{Total\ Number\ of\ Predictions}$$

For all experiments, we use the first 20 days in January, 2014 to train the model and the final 11 days to evaluate the performance of the model. Table II shows the results. We used bootstrapping to obtain the 95% confidence interval for each accuracy score. In our experiments, the Text-Enriched Model and the Text-Enriched with @-link Model produce statistically similar results. They both have better performance than the state-of-the-art baselines that we implemented.

TABLE II. Results for next-venue-type classification. Models with * are Baseline Models

Models	Prediction Accuracy (95% confidence interval)
Most Frequent Check-in Model* [8]	(0.5886,0.5915)
Order-1 Markov Model* [12]	(0.5575, 0.5758)
Order-2 Markov Model* [4]	(0.5282, 0.5661)
Classification model with historical visiting information*	(0.6333, 0.6476)
Text-enriched Model	(0.7122, 0.7158)
Text-enriched with @-link model	(0.7098, 0.7130)

TABLE III. Text-Retrieval Model Prediction Accuracy

Current Visiting Venue	Next Visiting Venue
0.1241	0.1321

The results in Table II support our hypothesis (H1) that the content of geotagged tweets correlates with users' future venue trajectories. The text-enriched model demonstrates better prediction performance than the classification model that uses historical venue trajectories. Although most users do not overtly mention their intent to move to a new venue type, they do reveal clues about this movement in the words of their messages. Extracting information pertaining to social relationships from @ mentions in tweets did not improve the text-enriched model, suggesting that people's movement plans may not be influenced by the textual content of messages posted by their network peers.

Table III shows the performance of the Text-Retrieval Model, which retrieves a venue type based on the similarity between the user's tweet and tweets posted from each venue type. We evaluated the accuracy of this model to determine whether the textual content of tweets correlates more with a user's current location or next location. We find that the textual content of users' tweets correlates more strongly with their next venue type instead of their current venue type. These results suggest that the Text-Enriched Model of next-venue classification might be improved in the future by incorporating a measure of similarity between a user's current tweet with tweets that have been posted from various venues (we leave this model for future work).

TABLE IV. Regression Model (MSE)

Venue Types	MSE (Baseline)	MSE (Main Effects)	MSE (interaction)
Transport	37365	10460	10262
Food	30433	12073	11795
Residence	26861	9831	9650
Recreation	30091	10665	10416
Professional	16127	8693	8450
Entertain	85713	18727	18443
Nightlife	40084	12482	12148
University	131126	26570	26404
Shop	20734	9233	9077
Event	745612	141189	140985

We used Mean Squared Error (MSE) as the performance metric for the venue-distance regression models. Table IV shows MSE results for the regression models on each of the 10 venue types in our data. We found that the University & College and Event venue types performed the worst, both in the baseline as well as the regression models. These venue types are typically concentrated in small areas of the city. As a result, tweets are often posted at great distances from these venue types, making accurate prediction a difficult task. However, both the main effects and interaction regression models showed improved performance in predicting the distances to venue types compared to the baseline model. These results support our hypothesis (H1), indicating that individuals' historical textual content correlates with the physical environment of his/her future trajectory. The regression model with interaction terms performs slightly better than the main effects model for each venue type. This suggests possible correlation between venue type distances, which is consistent with the intuition that certain venue types cluster together (e.g., food and shopping).

VI. Correlation Analysis of Crimes and Next-place Predictions

We hypothesized (H2) that crime counts would correlate with the predicted concentration of users at various venue types. This section presents analysis and results for this hypothesis.

A. Methodology

Users' movement patterns are defined as the predicted concentration of users at each venue type. When an individual posts one tweet, his/her movement pattern is the movement from the current location to the next location where a tweet is posted. We utilize the output of the next-place classifier to determine the concentration of users at their next venue types. For each venue type, we define the predicted occupants as follows:

$$Pre(p) = \{c_1(p), c_2(p), ...c_{10}(p)\}$$

where p is a spatial point in a grid of evenly spaced 2000-meter squares across the city, $c_i(p)$ is the count of predicted occupants for venue type i within the 2000-meter square that covers p. We performed these calculations at all points on every hour in January, 2015 in Chicago.

For crimes, we collected all records from Jan 1, 2014 - Jan 31, 2014 from the Chicago Data Portal. There are 25 crime types including 19,691 instances in total. Table V shows the crime count for each type in the city from January 2014. For the purposes of our study, we retained crime types with at least 1,000 instances. Similar to the next-place concentrations described above, we count the frequency of each crime type within a grid of 2000-meter squares that cover the city, on an hourly basis. These squares are identical to the squares used for next-place concentrations, allowing us to calculate basic correlation statistics as follows.

With the paired counts of crimes and users' movements to each venue type, we calculate the correlation between crime count and venue type occupancy. We used Pearson's Product Moment Correlation:

$$cor(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

TABLE V. Count of Crimes in Chicago, January 2014

Crime Type	Count
Gambling	1
Stalking	7
Intimidation	10
Kidnapping	20
Homicide	20
Arson	21
Liquor Law Violation	31
Sex Offense	65
Prostitute	68
Crime Sexual Assault	81
Interference with the Public Officer	90
Public Peace Violation	169
Weapon Violation	201
Offense Involving Children	230
Criminal Trespass	575
Robbery	797
Motor Cycle Violation	806
Burglary	1134
Assault	1036
Deceptive Practice	1138
Other Offense	1407
Criminal Damage	1789
Narcotics	2222
Battery	3335
Theft	4438

TABLE VI. Correlation Scores with Assault, * are with baseline features

Features	Correlation	p-value
Distance to Schools *	-0.032	0.000
Distance to Small Business *	-0.007	0.000
Arts and Entertainment	0.023	0.000
College and University	0.029	0.000
Event	-0.002	0.744
Food	0.026	0.000
Night Life Spot	0.011	0.048
Outdoors and Recreation	0.016	0.003
Professional and Other Places	0.035	0.000
Residence	0.024	0.000
Shop and Services	0.026	0.000
Travel and Transport	0.010	0.063

We calculated this correlation for ground-truth next-place trajectories as well as the predicted next-place trajectories from our Text-Enriched Classification Model. We estimated the exact p-value via the asymptotic t approximation to evaluate the significance of the correlation. The null hypothesis is that there is no correlation between crime count and venue type occupancy.

B. Results and Discussion

The results of our analysis are shown in Tables VI through XIV, where each table shows the correlation between crime

TABLE VII. Correlation Scores with Battery, * are with baseline features

Features	Correlation	p-value
Distance to Schools *	-0.055	0.000
Distance to Small Business *	-0.041	0.000
Arts and Entertainment	0.006	0.240
College and University	0.005	0.346
Event	-0.003	0.522
Food	0.017	0.002
Night Life Spot	0.009	0.093
Outdoors and Recreation	0.013	0.016
Professional and Other Places	0.030	0.000
Residence	0.012	0.027
Shop and Services	0.018	0.001
Travel and Transport	0.012	0.022

TABLE VIII. CORRELATION SCORES WITH BURGLARY, * ARE WITH BASELINE FEATURES

Features	Correlation	p-value
Distance to Schools *	-0.035	0.000
Distance to Small Business *	-0.033	0.000
Arts and Entertainment	0.005	0.318
College and University	-0.005	0.367
Event	-0.003	0.517
Food	0.001	0.813
Night Life Spot	0.005	0.388
Outdoors and Recreation	-0.004	0.472
Professional and Other Places	-0.004	0.418
Residence	0.007	0.170
Shop and Services	0.014	0.012
Travel and Transport	-0.002	0.659

TABLE IX. CORRELATION SCORES WITH CRIMINAL DAMAGE, * ARE WITH BASELINE FEATURES

Features	Correlation	p-value
Distance to Schools *	-0.040	0.000
Distance to Small Business *	-0.033	0.000
Arts and Entertainment	0.018	0.001
College and University	0.009	0.091
Event	0.001	0.822
Food	0.007	0.201
Night Life Spot	0.006	0.231
Outdoors and Recreation	0.002	0.748
Professional and Other Places	0.021	0.000
Residence	0.012	0.068
Shop and Services	0.013	0.000
Travel and Transport	0.007	0.306

counts and predicted concentrations of users at the various venue types. Across the tables, we see that many crime types have significant correlations with predicted venue occupancy concentrations (indicated by p-values). The significant correlations support our hypothesis that crime is correlated with transition among venue types. For example, the occurrence of burglaries (Table VIII) is positively correlated with the transition of users to Shop & Services destinations. At this point in our work, we are not clear on the causal mechanism that underlies this correlation; however, it is consistent with the intuition that burglaries are prevalent in places where residents have left their homes, e.g., to travel to shopping or service centers. We also noticed that transitions to Residence and Professional & Other Places are informative in almost all types of crimes, suggesting an increase in general activity is correlated with crime count. However, the best approach for integrating these features into a full crime prediction system [15] remains an open question for future work.

From Tables VI - XIV, we should also note that most correlation scores are positive. These results are consistent with the hypothesis that mere concentration of individuals, regardless of the venue type, increases risk of crime. The only exception to this observation is Narcotics (Table XI) and its negative correlation with transitions to College and University venues. One possible explanation for this is the increased security presence at such venues, which could deter such activity.

Based on our results, we also note that certain venue types exhibited similar correlation scores. For example, Food and Night Light Spot venue types show similar correlations across many crime types. One possible explanation for this is the spatial correlation of venue types: Food and Night Life Spots are often located near each other.

TABLE X. CORRELATION SCORES WITH DECEPTIVE PRACTICE, * ARE WITH BASELINE FEATURES

Features	Correlation	p-value
Distance to Schools *	-0.028	0.000
Distance to Small Business *	-0.037	0.000
Arts and Entertainment	0.063	0.000
College and University	0.114	0.000
Event	0.000	0.948
Food	0.048	0.000
Night Life Spot	0.032	0.000
Outdoors and Recreation	0.047	0.000
Professional and Other Places	0.095	0.000
Residence	0.037	0.000
Shop and Services	0.053	0.000
Travel and Transport	0.064	0.000

TABLE XI. CORRELATION SCORES WITH NARCOTICS, * ARE WITH BASELINE FEATURES

Features	Correlation	p-value
Distance to Schools *	-0.054	0.000
Distance to Small Business *	-0.029	0.000
Arts and Entertainment	0.006	0.261
College and University	-0.012	0.024
Event	-0.004	0.483
Food	0.013	0.017
Night Life Spot	0.006	0.286
Outdoors and Recreation	0.006	0.239
Professional and Other Places	.039	0.000
Residence	0.010	0.068
Shop and Services	0.025	0.000
Travel and Transport	-0.006	0.306

TABLE XII. CORRELATION SCORES WITH THEFT, * ARE WITH BASELINE FEATURES

Features	Correlation	p-value
Distance to Schools *	-0.058	0.000
Distance to Small Business *	-0.069	0.000
Arts and Entertainment	0.058	0.000
College and University	0.106	0.000
Event	0.000	0.981
Food	0.066	0.000
Night Life Spot	0.059	0.000
Outdoors and Recreation	0.050	0.000
Professional and Other Places	0.112	0.000
Residence	0.047	0.000
Shop and Services	0.072	0.000
Travel and Transport	0.064	0.000

VII. TOWARD A BETTER CRIME PREDICTION MODEL

Having demonstrated an ability to predict the types of next venues as well as statistically significant correlations between next-place occupancy patterns and crime rates, we sought to move closer to a full crime prediction model such as the one developed by [15]. We laid points down across the

TABLE XIII. CORRELATION SCORES WITH OTHER OFFENSE, * ARE WITH BASELINE FEATURES

Features	Correlation	p-value
Distance to Schools *	-0.020	0.000
Distance to Small Business *	-0.027	0.000
Arts and Entertainment	0.011	0.044
College and University	0.007	0.176
Event	-0.001	0.830
Food	0.012	0.021
Night Life Spot	0.001	0.875
Outdoors and Recreation	0.006	0.239
Professional and Other Places	0.024	0.000
Residence	0.024	0.000
Shop and Services	0.008	0.155
Travel and Transport	0.008	0.152

TABLE XIV. CRIME PREDICTION WITH NEXT VISITING VENUE COUNT

Result	All Crimes
Accuracy	0.3456
Precision	0.1623
Recall	0.8006
F1	0.1349

city boundary of Chicago, evenly spaced points at 2000-meter intervals. The label for point p is $true$ if, in the past one hour, a crime occurred in the square covering p; otherwise, the label for point p is $false$. Thus, we established a binary classification problem for differentiating areas with crime from those without crime on an hourly basis. The general form of this classification problem is as follows:

$$Label(p)|c_1(p), c_2(p), ...c_{10}(p)$$

where features $c_1(p), c_2(p), ...c_{10}(p)$ come from our next-place prediction models and represent concentrations of users at the 10 venue types. We used support vector machines to optimize the weights of the venue types [7].

Table XIV shows classification performance for the SVM described above. As shown, by only examining the predicted concentrations of individuals at the 10 venue types, the SVM classifier is able to differentiate crime from non-crime points with an accuracy of 35%. Thus, we believe that overall crime prediction performance might be improved by incorporating $c_1(p), c_2(p), ...c_{10}(p)$ into a fuller model of crime that controls for many other spatiotemporal explanatory variables [31].

VIII. CONCLUSION AND FUTURE WORK

Our experimental results support our hypotheses: We find that the textual content of tweets improves next-place prediction compared with baselines that do not consider tweets' textual content. We further find evidence of correlation between predicted next-place concentrations at various venue types and the occurrence of crime.

Future work should consider a few major aspects of these problems. It would be interesting to give further consideration to the network of relationships present in the Twitter data. In our experiments, we did not find benefit in using the @-link information for next-place prediction, but this was just a preliminary model and we believe that further investigation might uncover interesting correlations between the venue trajectories of users' friends and the users' themselves. In the venue distance regression setting, one might expect to observe correlations between distances to certain venue types. For example, one might expect to see restaurants near arts & entertainment venues. Thus, the prediction of a user's future distances to such venue types should be constrained to take such correlations into account. Regarding crime, we have found preliminary evidence of correlation between predicted next-places and the occurrence of crime; however, we have yet to incorporate these correlations into a full crime prediction model, such as the ones described by [15].

REFERENCES

[1] D. Ashbrook and T. Starner. Learning significant locations and predicting user movement with gps. In *Wearable Computers, 2002.(ISWC 2002). Proceedings. Sixth International Symposium on*, pages 101–108. IEEE, 2002.

[2] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[3] S. J. Barnes and E. Scornavacca. Mobile marketing: the role of permission and acceptance. *International Journal of Mobile Communications*, 2(2):128–139, 2004.

[4] P. Baumann, W. Kleiminger, and S. Santini. The influence of temporal and spatial features on the performance of next-place prediction algorithms. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 449–458. ACM, 2013.

[5] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. " O'Reilly Media, Inc.", 2009.

[6] S. Chainey, L. Tompson, and S. Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1):4–28, 2008.

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[8] J. Chang and E. Sun. Location 3: How users share and respond to location-based data on social networking sites. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 74–80, 2011.

[9] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[10] T. M. T. Do and D. Gatica-Perez. Contextual conditional models for smartphone-based human mobility prediction. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 163–172. ACM, 2012.

[11] T. M. T. Do and D. Gatica-Perez. Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing*, 12:79–91, 2014.

[12] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, page 3. ACM, 2012.

[13] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *ICWSM*, 2012.

[14] H. Gao, J. Tang, and H. Liu. Mobile location prediction in spatio-temporal context. In *Nokia mobile data challenge workshop*. Citeseer, 2012.

[15] M. S. Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[16] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics, 2011.

[17] M. F. Goodchild and J. A. Glennon. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3):231–241, 2010.

[18] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. ACM, 2011.

[19] D. Jiang, X. Guo, Y. Gao, J. Liu, H. Li, and J. Cheng. Locations recommendation based on check-in data from location-based social network. In *Geoinformatics (GeoInformatics), 2014 22nd International Conference on*, pages 1–4. IEEE, 2014.

[20] D. Lian, V. W. Zheng, and X. Xie. Collaborative filtering meets next check-in location prediction. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 231–232. International World Wide Web Conferences Steering Committee, 2013.

[21] Z. Lu, Y. Zhu, V. W. Zheng, and Q. Yang. Next place prediction by learning with multiple models.

[22] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[23] F. Miró. Routine activity theory. *The Encyclopedia of Theoretical Criminology*, 2014.

[24] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 637–646. ACM, 2009.

[25] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In *ICDM*, volume 12, pages 1038–1043. Citeseer, 2012.

[26] S. Robertson and H. Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[27] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell. Nextplace: a spatio-temporal prediction framework for pervasive systems. In *Pervasive Computing*, pages 152–169. Springer, 2011.

[28] L. W. Sherman, P. R. Gartin, and M. E. Buerger. Hot spots of predatory crime: Routine activities and the criminology of place*. *Criminology*, 27(1):27–56, 1989.

[29] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[30] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1079–1088. ACM, 2010.

[31] X. Wang and D. Brown. The spatio-temporal generalized additive model for criminal incidents. In *Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on*, pages 42–47. IEEE, 2011.

[32] X. Wang, M. S. Gerber, and D. E. Brown. Automatic crime prediction using events extracted from twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 231–238. Springer, 2012.