CrimeTelescope: crime hotspot prediction based on urban and social media data fusion

Dingqi Yang[1] · **Terence Heaney[1]** · **Alberto Tonon[1]** ·
Leye Wang[2] · **Philippe Cudré-Mauroux[1]**

Abstract Crime is a complex social issue impacting a considerable number of individuals within a society. Preventing and reducing crime is a top priority in many countries. Given limited policing and crime reduction resources, it is often crucial to identify effective strategies to deploy the available resources. Towards this goal, crime hotspot prediction has previously been suggested. Crime hotspot prediction leverages past data in order to identify geographical areas susceptible of hosting crimes in the future. However, most of the existing techniques in crime hotspot prediction solely use historical crime records to identify crime hotspots, while ignoring the predictive power of other data such as urban or social media data. In this paper, we propose CrimeTelescope, a platform that predicts and visualizes crime hotspots based on a fusion of different data types. Our platform continuously collects crime data as well as urban and social media data on the Web. It then extracts key features from the collected data based on both statistical and linguistic analysis. Finally, it identifies crime hotspots by leveraging the extracted features, and offers visualizations of the hotspots on an interactive map. Based on real-world data collected from New York City,

✉ Dingqi Yang
Dingqi.Yang@unifr.ch

Terence Heaney
Terence.Heaney@unifr.ch

Alberto Tonon
Alberto.Tonon@unifr.ch

Leye Wang
wly@cse.ust.hk

Philippe Cudré-Mauroux
Philippe.Cudre-Mauroux@unifr.ch

[1] eXascale Infolab, Department of Computer Science, University of Fribourg, Bd de Perolles 90, 1700 Fribourg, Switzerland

[2] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

we show that combining different types of data can effectively improve the crime hotspot prediction accuracy (by up to 5.2%), compared to classical approaches based on historical crime records only. In addition, we demonstrate the usability of our platform through a System Usability Scale (SUS) survey on a full prototype of CrimeTelescope.

Keywords Crime prediction · Data fusion · Urban open data · Social media

1 Introduction

Crime, in all its facets, has been recognized as an important social problem for public safety. Being able to identify risky places for crime has become an increasing concern for both urban authorities and citizens [35]. On one hand, urban authorities are always interested in the improvement of public safety in urban neighborhoods. Given limited policing and crime reduction resources, it is important to find optimal deployment strategies. On the other hand, crime rate is considered as an essential factor that determines whether or not people move to a new neighborhood or what places should be avoided when traveling [1].

In criminological studies, crime hotspot map is a popular analytical tool to discover and visualize the areas of high concentrations of crime (i.e., crime hotspots) [11]. In practice, they are widely used as a basic form of crime prediction, relying on retrospective criminal activity data to identify whether a geographical area is likely to be a crime scene or not [6, 11, 16]. Such a crime hotspot map would be beneficial not only for urban authorities when taking decisions on where to deploy policing and other crime reduction resources, but also for citizens when examining the public safety in their neighborhoods and in the places they plan to travel. Therefore, it is of critical importance for both urban authorities and citizens to have a crime hotspot map with high predictive ability that accurately reflects the crime risk across places.

Currently, existing systems for crime hotspot maps—including both open data portals (such as NYC Open Data[1]) and systems designed specifically for analyzing criminal activities (such as Crimespotting[2], SpotCrime[3] and CrimeMapping[4])—collect crime data from the local police department and visualize them on an interactive map. However, these systems are built based on historical crime records only, resulting in limited predictive ability for the generated crime hotspot maps. Environmental criminology suggests putting focus of criminological study on environmental factors that can influence criminal activities [8]. Following this direction, researchers have been focusing on analyzing the correlations between the aggregated crime activities and various urban socio-economic statistics across different geographical areas (including demographics [6], education [12], unemployment [26], ethnicity [24] and urban infrastructures [18], etc.), and leveraging such correlations to improve crime hotspot prediction accuracy.

However, the fact that those socio-economic statistics are often not collected in a timely manner prevents the accurate prediction of crime hotspots. For example, demographic data in the United States are mostly collected and reported on a yearly basis[5]. Given the large

---

[1]https://nycopendata.socrata.com/

[2]http://stamen.com/work/crimespotting/

[3]https://www.spotcrime.com/

[4]http://www.crimemapping.com/

[5]https://www.census.gov/ces/dataproducts/demographicdata.html

number of floating population in urban areas, such annual demographic data become less effective over time for crime prediction. Fortunately, with the recent popularity of urban Big Data [19], large collections of urban and social media data has become available, which are often accessible in a streaming manner and can be easily maintained up-to-date. More importantly, those data often characterize the urban environment in different aspects. For example, urban elements can be represented by Points of Interests (POIs) [46], such as bars, supermarkets or universities; demographic attributes are often linked to user generated content on social networks such as Twitter [47]. Therefore, such urban and social media data has great potential to improve the accuracy of crime hotspot prediction.

In this paper, we propose CrimeTelescope, a platform that predicts and visualizes crime hotspots based on the fusion of heterogeneous urban and social media data. Specifically, CrimeTelescope is designed to integrate various urban and social media data with historical crime data to provide a high-quality and user-friendly crime hotspot prediction system. To achieve this goal, it first continuously collects historical crime data and two types of urban and social media data including urban infrastructures data and Tweets. Second, it extracts features from the above data by applying statistical and linguistic analysis. Finally, it predicts crime hotspots using classification techniques. In addition, CrimeTelescope is designed with an interactive Web interface to visualize the predicted crime hotspots on a map. Our contributions can be summarized as follows:

– To the best of our knowledge, CrimeTelescope is the first online system predicting crime hotspots using urban and social media data fusion.
– To deliver a higher-quality crime hotspot map, we propose an efficient crime hotspot prediction approach by considering various urban and social media data. Specifically, the proposed approach can extract representative features from various data sources and integrate them for crime hotspot prediction.
– To provide a user-friendly interface to end users, we design a simple yet informative Web interface to visualize the predicted crime hotspots on an interactive map.
– CrimeTelescope is evaluated from both an effectiveness (crime prediction performance) and a system usability perspective. The results show that it can effectively predict crime hotspots and offers excellent usability. Particularly, compared to classical approaches based on historical crime records only, CrimeTelescope is able to improve the prediction accuracy by up to 5.2%.

The rest of the article is organized as follows. Section 2 presents related work. Sections 3 and 4 illustrate the design of the CrimeTelescope platform and give details on our crime hotspot prediction process, respectively. Section 5 presents the implemented prototype of CrimeTelescope and its functionalities. Sections 6 and 7 present our empirical evaluation, focusing on crime prediction performance and system usability, respectively. We conclude our work in Section 8.

2 Related work

As an inevitable and important social problem of our societies, criminal activities have been traditionally studied in different fields, such as sociology [15], psychology [13], and economics [23]. The recent movements of open data have made large repositories of crime activity data become publicly accessible, which provides an unprecedented opportunity to start studying crime in a more systematic manner. Along with the recent advances in Big Data analytic techniques, we are able to extensively analyze and deeply understand crime

3

patterns. In the current literature, researchers have devoted their attention on analyzing criminal activities mainly from two different perspectives, i.e., people-centric and place-centric perspectives.

On one hand, studies from a people-centric perspective mainly focus on discovering individual/collective criminal profiles. For example, Wang et al. [39, 40] have recently conducted a series of work on applying machine learning techniques to detect specific patterns of criminal activities committed by the same offender (or group of offenders). Although the people-centric studies reveal interesting patterns of criminal activity, they often do not consider the environmental or context factors that influence criminal activities, which is the main focus of environmental criminology [8].

On the other hand, studies from a place-centric perspective try to understand the geographical topology of criminal activities. Specifically, it has been widely shown that criminal activities often exhibit complex relationships both in space and time [25, 33]. Against this background, hotspot maps are a popular tool for analyzing and visualizing the distribution of crimes across space and time [11], and are also practically used for crime prediction in order to optimize the deployment of the police and crime reduction resources. For example, real-life experiments have shown that place-centric policing can significantly reduce the violent crime incidents by 23% in Philadelphia [27], 14% in Boston [7] and 33% in Jacksonville [32].

In this paper, we advocate for a place-centric study of crime activities. As suggested by environmental criminology, there is a strong influence on the criminal activities from the corresponding environmental or context factors. Traditionally, researchers have investigated the correlation between criminal activities and various socio-economic factors including demographics [6], education [12], unemployment [26], ethnicity [24] and urban infrastructures [18]. Although such correlation can be leveraged to improve crime prediction, it falls short in rendering an up-to-date crime hotspot map, as those socio-economic statistics are not often collected in a timely manner (often on a yearly basis).

The recent popularity of urban and social media data brings an unprecedented opportunity for criminological studies. The advantage lies on that such data can be easily accessible in a realtime manner and, more importantly, on the fact that it is correlated with the criminal activities. For example, it has been shown that local crime rate is actually correlated with the surrounding POIs, which can be collected from location based social network [38]; it is also shown that the criminal activities of certain types in a neighborhood are linked to the Tweets posted from there [16]. Motivated by such findings, we study in this paper the problem of crime hotspot prediction based on heterogeneous urban and social media data fusion. Different from the existing work that mainly use single and static datasets, we focus on leveraging data fusion from multiple data sources to develop an online system of crime prediction. Moreover, in addition to the classical quantitative evaluation on crime prediction based on real-world data, we also conduct a system usability scale survey to validate our system.

3 Platform design

In this section, we present the platform design of CrimeTelescope. Figure 1 shows its system architecture and the end-to-end workflow (from data sources to Web interface). Specifically, CrimeTelescope mainly consists of five parts, viz., Data Collector, Feature Extractor, Crime Predictor and Crime Hotspot Visualizer, as well as a Main Database. First, open crime data and social media data are collected by the Data Collector and get stored in
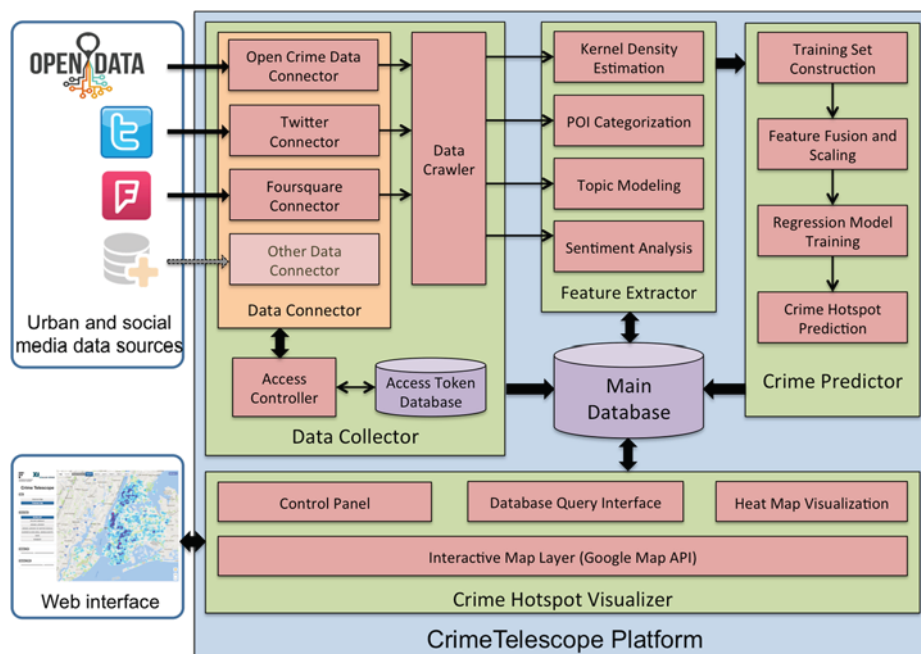
4

Figure 1 System architecture of the CrimeTelescope platform

the Main Database. Second, the Feature Extractor processes and mines the collected urban and social media data to extract features. Third, by merging the extracted features, Crime Predictor learns a classification model for crime prediction, and then stores results in the Main Database. Finally, the Data Visualizer supports a Web based interactive map interface for visualizing crime hotspots. In the following, we present the detailed design and characteristics of each component.

3.1 Data collector

The Data Collector is responsible for collecting related urban and social media data from open data portal and social network service providers. As illustrated in Figure 1, it is composed of several Data Connectors, a Data Crawler, an Access Controller and an Access Token Database. Due to security considerations, most open data portals and social network services have integrated the OAuth protocol[6], an open standard for authorization. In order to access the data, an authentication process is required with data-provider-specific access tokens. To achieve this goal, the Access Controller and Access Token Database is implemented for authentication with different open data portals and social networks. Moreover, since each individual service provider usually has its own specific Application Programming Interfaces (APIs), the corresponding Data connector is implemented under such specifications. After the authentication, the Data Crawler collects urban and social media data. Such

---

[6]http://oauth.net/

a design of the Data Collector has advantages on its extensibility, i.e., new data sources can be easily incorporated by only implementing the corresponding connectors.

3.2 Feature extractor

The Feature Extractor component is responsible for data preprocessing, analysis and feature extraction. As shown in Figure 1, for each specific type of data, it adopts the corresponding processing techniques for feature extraction. Specifically, it uses Kernel Density Estimation on crime data, POI categorization on urban infrastructure data, topic modeling and sentiment analysis techniques on Tweets. The feature extraction will be presented in more detail below in Section 4.

3.3 Crime predictor

The Crime Predictor component is responsible for generating a high-quality crime hotspot map by combining the features extracted from various data sources. Here, we formulate the crime prediction problem as a classification problem, where we aspire to classify whether a given location is a crime scene or not. To achieve this goal, the Crime Predictor merges and rescales the extracted features, and then learns a classification model. The learned model outputs the predicted probability of a given location being a crime scene. Finally, this predicted crime probability is stored in the Main Database.

3.4 Crime hotspot visualizer

The Crime Hotspot Visualizer is responsible for rendering the crime hotspot visualization on an interactive map. As presented in Figure 1, it consists of four components. First, the Database Query Interface provides the access to the Main Database. Then, the crime hotspots are visualized based on a heat map, where a control panel is provided with multiple options on both data and visual effects. The data visualization is built upon an interactive map layer, which is implemented using Google Maps API[7].

4 Crime hotspot prediction

In this section, we present the crime hotspot prediction process implemented in the CrimeTelescope platform. Specifically, by formulating the crime prediction problem as a classification problem, we combine features extracted from multiple data sources, and then feed them into a classification model to predict the probability of a given location being a crime scene. In the following, we first present an overview of our approach, followed by the data collection and feature extraction, and by the crime prediction process.

4.1 Approach overview

To predict crime from a place-centric perspective, we formulate the following classification task: given a geo-location l and its feature vector F_l, we train a classification model to output the probability of l being a crime scene. Here, the feature vector is extracted from

---

[7]https://developers.google.com/maps/

6

various social and urban data, which will be presented in Section 4.2. In order to construct a training dataset, we consider historical crime incidents as positive samples. However, as historical crime data contains only information about crime incidents, there is no negative sample (i.e., "non-crime incidents") available. To solve this problem, we randomly and uniformly generate negative samples from the map as suggested by [16]. More precisely, we sample evenly spaced points, which are not coinciding with the positive samples, and regard these points as negative samples. The sampling granularity actually controls the number of negative samples. In order to avoid getting an imbalanced class distribution (i.e., a biased ratio of positive/negative samples), which brings a serious difficulty to most classification algorithms [31], we tune the sampling granularity such that the number of negative samples is similar to that of positive samples. Subsequently, we train a classification model based on the constructed training data, which can then evaluate the crime probability of any location l. More information on the prediction method will be given in Section 4.3.

4.2 Data collection and feature extraction

In this study, we focus on a use case in New York City (NYC), as its open data portal is well developed. However, our approach and platform are not limited to any specific city. New York City is composed of five boroughs i.e., Manhattan, the Bronx, Queens, Brooklyn, and Staten Island. Figure 2 shows the map of New York City with its five boroughs. The Data Crawler collects the three types of data (i.e., crime data, urban infrastructure data and Tweet data) within the NYC boundaries.

4.2.1 Crime data

Historical crime records have been regarded as the primary data source for crime prediction [6, 11, 16]. Many studies in criminology have provided evidence of significant concentration of crime at a geographical level (i.e., crime hotspots), and found that future crimes often occur in the vicinity of those areas [9, 16]. Therefore, we collect historical crime data from the NYC open data portal[8]. It contains incidents of 7 major crime types (i.e., "burglary", "felony assault", "grand larceny", "grand larceny of motor vehicle", "murder & non-negl manslaughte", "rape" and "robbery"). Each incident is recorded with detailed information about its time, location (including GPS coordinates) and crime type. As the data is periodically updated on the open data portal, we update our historical crime dataset accordingly. To give a statistical snapshot, 7395 incidents were reported in March 2015. Tables 1 and 2 show the related statistics of the crime data in different boroughs and crime types, respectively.

In order to predict crime based on historical data, Kernel Density Estimation (KDE) has been widely adopted to fit a two-dimensional spatial probability density function to historical crime records. The resulted probability density function has a higher value for locations with more offenses close to the selected location point, or a lower value if there are fewer offenses around the selected location point. Such a probability density function indicates the potential risk of crimes at different locations in a city, which was shown as very effective for prediction [11].

In this study, we advocate such a method and consider the fitted probability density estimates as the feature extracted from the historical crime data. Specifically, for a
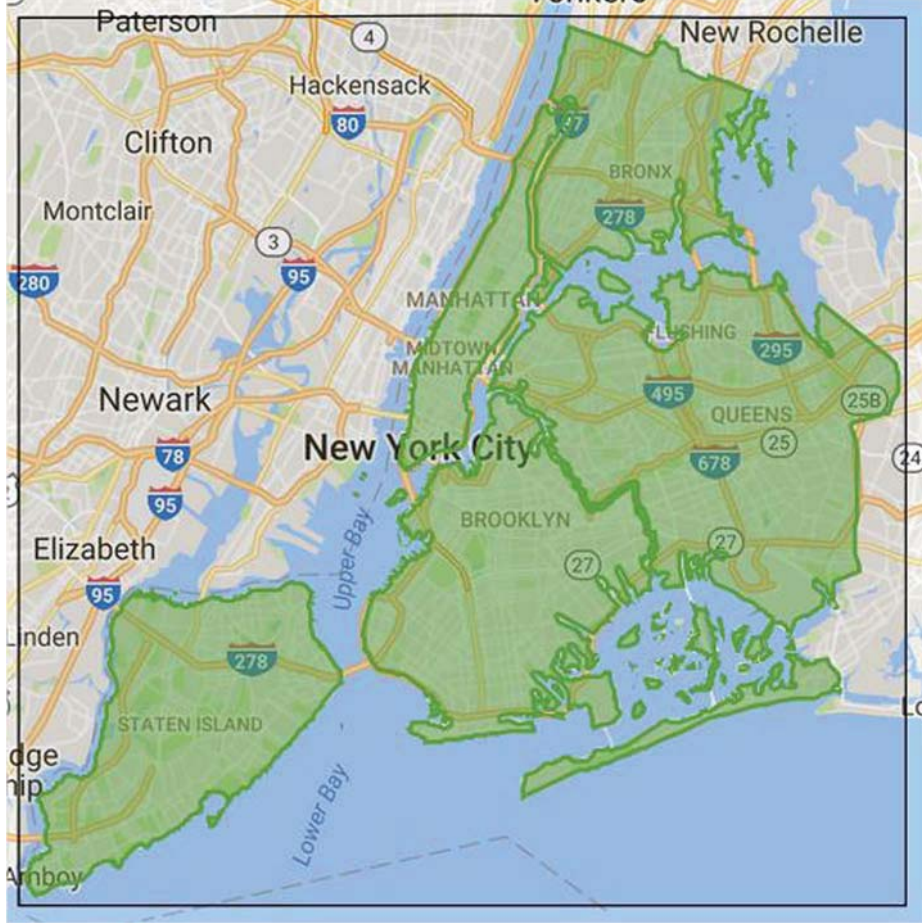
---

[8]https://nycopendata.socrata.com/

7

Figure 2 New York City boundary map

given location $l = (x, y)$, (where x and y are latitude and longitude, respectively), the two-dimensional kernel density estimate [28] is defined as:

$$\hat{f}_H(l) = \frac{1}{n} \sum_{i=1}^{n} |H|^{-1/2} K(H^{-1/2}(l - l_i)) \tag{1}$$

where n is the number of crime incidents; H is a bandwidth matrix of size 2×2; K is the kernel function. In this study, we select K as Gaussian kernel, as it has shown a good performance in estimating hotspots over space [16]. We set the optimal bandwidth matrix H using Scott's rule [28], which is a commonly used approach to select the bandwidth:

$$H = n^{-1/6} cov(x, y) \tag{2}$$

where $cov(x, y)$ is the covariance matrix of x and y. Please refer to [28] for more details. Subsequently, we can calculate the density estimate $\hat{f}_H(l)$ and use it as a feature for location l. We formally denote F^c the feature extracted from crime data, i.e., $F_l^c = \hat{f}_H(l)$.

8

Table 1 Statistics of crime data in different boroughs (March 2015 in NYC)

Borough	Incident number	Percentage
Queens	1,588	21.47%
Brooklyn	2,243	30.33%
Manhattan	1,902	25.72%
Bronx	1,423	19.24%
Staten Island	239	3.23%

4.2.2 Urban infrastructure data

Criminological studies have recognized urban infrastructure as an important factor influencing criminal activities. For example, it has been empirically verified that bars/nightclubs hosting social events will increase the probability of disturbances, crime, and violence [5]. In this study, we characterize urban infrastructure using POIs. Specifically, we collect POI data in New York City from Foursquare[9]. Each POI is associated with its name, location and category. We focus on POI categories as they represent urban infrastructure [46] (e.g., bars, museums or colleges). More precisely, Foursquare classifies POIs using a three-level hierarchical category classification[10]. By March 2015, we collected 67,265 POIs of 9 root categories (i.e. Arts & Entertainment, College & University, Food, Great Outdoors, Nightlife Spot, Professional & Other Places, Residence, Shop & Service, Travel & Transport), which are further classified into 291 categories at the second level. Moreover, a few second-level categories have sub-categories at the third level. We regularly query Foursquare to maintain an up-to-date POI dataset.

To characterize urban infrastructures from POI data, the urban computing community often resorts to the categorical distribution of POIs [45, 46]. Motivated by this idea, we also consider the categorical distribution of POIs as the features for crime prediction. Due to the incompleteness of third-level categories, only a few POIs are associated to a third-level category. Therefore, we focus on the categorical distribution of POIs at the first and second levels. Table 3 shows the overall statistics of POI categories at the first level. As the number of second level POI categories is large (291 categories), we plot a word cloud to visualize their distribution in Figure 3.

For the purpose of crime prediction, we need to extract features for a given location l from its surrounding POIs. Therefore, we first split the NYC territory into small grid cells, and then regard the categorical distribution of POIs in each cell as the feature of the locations there. Specifically, we generate a grid of cells on a 0.01 latitude by 0.01 longitude resolution. Figure 4 shows the resulting grid cells. For a given location l, we first find the cell that it locates in, and consider the categorical distribution of POIs in that cell as the feature extracted for location l. We denote such urban infrastructure feature as F_l^p. Figure 5 shows the distribution of the second-level POI categories for two different grid cells. As can be seen, we find obvious differences between the urban infrastructure there. We observe a lot of offices around Wall street, while Time square is surrounded by many restaurants and shops.

---

[9]https://foursquare.com/

[10]https://developer.foursquare.com/categorytree

Table 2 Statistics of crime data in different types (March 2015 in NYC)

Crime type	Incident number	Percentage
Burglary	1,031	13.94%
Felony assault	1,476	19.96%
Grand larceny	3,087	41.74%
Grand larceny of motor vehicle	511	6.91%
Murder & non-negl. manslaughte	23	0.31%
Rape	94	1.27%
Robbery	1,173	15.86%

Table 3 Statistics of POI categories at first level in NYC

POI category	POI number	Percentage
Arts & entertainment	3,529	5.24%
Shop & service	14,315	21.28%
Food	18,134	26.95%
Nightlife spot	4,111	6.11%
Professional & other places	10,160	15.10%
Travel & transport	6,247	9.28%
Outdoors & recreation	4,581	6.81%
College & university	1,686	2.50%
Residence	4,502	6.69%

Figure 3 Distribution of second-level POI categories in NYC (Plotted as a word cloud. Larger font size implies higher frequency, and vice versa)

Figure 4 New York City grid cells (on a 0.01 latitude by 0.01 longitude resolution)

4.2.3 Tweet data

Twitter[11] users can share a short message with their current location in a realtime manner. Such content has been leveraged in many predictive applications, such as election prediction [36], disease spreading forecasting [30] and critical event detection [37]. The essential idea of those work is that the content of Tweets is informative with regard to future events. Following this idea, researchers have recently studied how to use Tweets to tackle the crime prediction problem [16, 41]. The basic intuition is that the information encoded by Tweets are somehow correlated with crime rate. For example, it has been empirically verified that social events taking place at nightclubs will increase the probability of disturbances, crime, and violence [5]. Therefore, we also incorporate Tweet data in our platform. We continuously collect Tweets in New York City from the Twitter Public Streams[12] in an online manner. As a statistical snapshot, we collected 750,369 Tweets in March 2015.

We extract two types of features from Tweets, namely topic feature and sentiment feature. The topic feature characterizes what people talk about in an area by a distribution of potential topics, while sentiment feature shows the distribution of sentiment polarity in the area.

For the topic feature, we adopt topic modeling techniques (Latent Dirichlet Allocation (LDA) [4]), which are also used in [16]. The processing workflow is shown in Figure 6. First, we perform the language detection to filter out non-English Tweets using the language detection library developed by Cybozu Labs [29]. Second, we aggregate Tweets by mapping them onto grid cells on the map. Here we consider the same grid cells as shown in Figure 4.

---

[11]https://twitter.com/

[12]https://dev.twitter.com/streaming/public

(a) A grid cell around Wall street

(b) A grid cell around Time Square

Figure 5 Distribution of POI categories in two different grid cells

For each grid cell, we aggregate all Tweets within a certain time period that are reported with a location in that grid cell, and consider the aggregated content as a document. (We will later test different time periods for selecting the optimal one.) Third, we perform a Tweet-specific tokenization from NLTK toolkit [3] on the documents. The advantages of using Tweet-specific tokenization mainly lie on that it preserves emoticons (e.g., ";-)") in the Tweets. Fourth, we filter out stop words including both common English stop words (e.g., "the") and our self-defined stop words [13] (e.g., "new" and "york"). Finally, by considering the documents from all the grid cells as a corpus, we apply Latent Dirichlet Allocation (LDA) [4] to train a topic model. Such a topic model can output a distribution of topics for each grid cell, where each topic is a distribution of words describing the topic. For example, a grid cell around John F. Kennedy airport may have a topic distribution with a high probability on the "airport" topic, which contains keywords such as "gate", "flight", "airlines", "delayed", "luggage", etc. In this study, we empirically set the number of topics to 50. By characterizing each grid cell using its distribution on topics, we obtain its topic feature, denoted by F_l^{t1}.

For the sentiment feature, we analyze individual Tweets using text-based sentiment analysis techniques [42]. As shown in Figure 6, we first filter out non-English Tweets, and then use the Sentiment140 API [17] to obtain the sentiment polarity for individual Tweets. Sentiment140 defines three classes of sentiment polarity, i.e., negative, neutral and positive. It classifies each Tweet to one sentiment polarity class using distant supervision. Please refer to [17] for more details. After obtaining the sentiment polarity for individual Tweets, we aggregate the sentiment for each grid cell on the map by empirically computing its distribution on the three sentiment polarity classes. We thus consider such sentiment distribution as sentiment feature, denoted by F_l^{t2}.

Based on the extracted topic and sentiment features, we merge them to obtain the overall feature extracted from Tweets F_l^t by concatenating them together $[F_l^{t1}, F_l^{t2}]$.

4.3 Problem formulation of crime prediction

In this section, we formally define our crime prediction problem. Specifically, the objective of the crime prediction in this paper is to predict if a location will be a crime scene or not,

---

[13]Note that the self-defined stop words are iteratively selected. More precisely, analyzing the LDA results allows us to detect frequent words that do not add any meaning to the documents and include them in the stop word list for another round of LDA training.
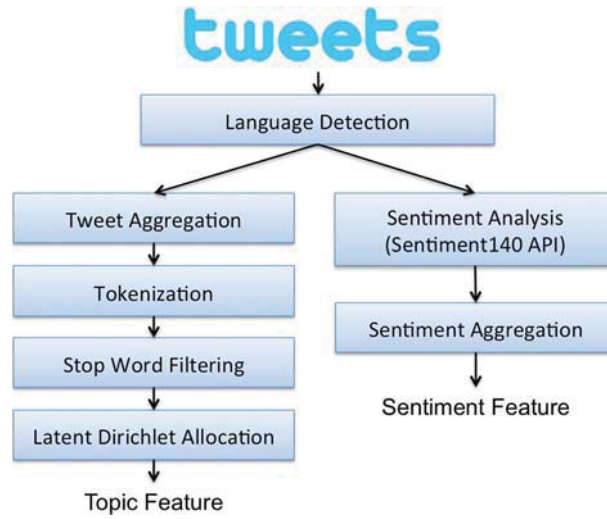
Figure 6 Topic and sentiment feature extraction from Tweets

which is intrinsically a binary classification problem. Given a location l, following the above feature extraction process, we can get three sets of features F_l^c, F_l^p, F_l^t from crime data, urban infrastructure data and Tweet data, respectively. By concatenating these features, we obtain the concatenated feature vector $F_l = [F_l^c, F_l^p, F_l^t]$ for location l. The problem now is to classify whether a location l will be a crime scene or not based on the feature vector F_l.

However, two issues need to be solved when defining this problem. First, as a binary classification task requires that the training data contains both positive (i.e., crime scene) data samples and negative (non-crime scene) data samples, we have only historical crime incidents representing positive data samples. Therefore, we need to design an appropriate approach to get the negative samples over space. Second, due to the dynamics of criminal activities, a learned crime prediction model will become less and less effective over time. Therefore, we need to define a mechanism to maintain the effectiveness of our crime prediction model. In the following, we propose two solutions to address these two issues, respectively.

4.3.1 Negative data sampling

In order to get negative samples, we use the sampling method proposed in [16]. Its basic idea is to generate evenly-spaced locations at a certain sampling granularity, and consider those locations that are not coinciding with the positive samples as negative samples. Figure 7 shows an example of generating negative samples. Given the positive data samples (locations of crime scenes) shown as green points, we firstly generate evenly-spaced locations with a certain distance interval, and then remove the ones that are coinciding with positive samples, resulting in the negative samples shown as red points. The coinciding points are empirically defined as points that are less than 200 meters from any crime scenes. The sampling distance interval actually controls the number of negative samples.

Moreover, in this study, we further consider the crime prediction problem for individual crime types, since existing criminological studies have shown that different crime types
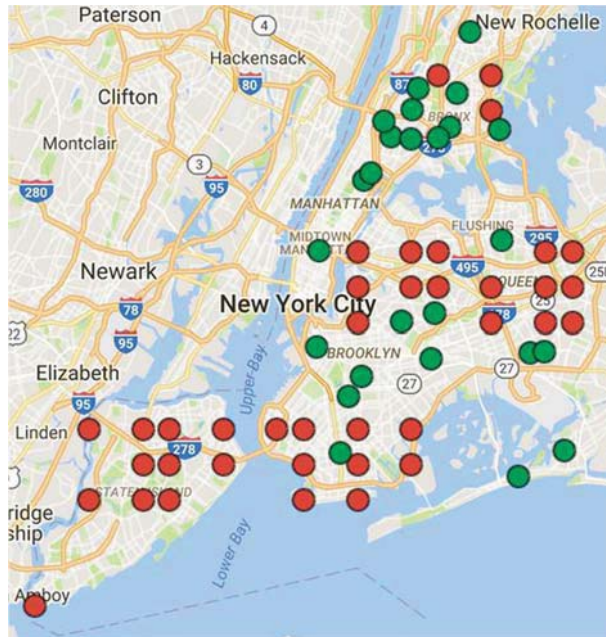
Figure 7 An example of generating negative data samples (Green points refer to the positive data samples (crime scenes), while red points refer to the sampled evenly-spaced negative data points)

exhibit different patterns of hotspots [16]. Therefore, we build a training dataset for each crime type. Particularly, as the number of crime incidents varies across different crime types, we tune the sampling distance interval to obtain a balanced number of negative samples, i.e., the number of negative samples N_{neg} is within $[0.8N_{pos}, 1.2N_{pos}]$, where N_{pos} is the number of positive samples in the historical crime data.

4.3.2 Update frequency of crime prediction model

To evaluate our crime prediction model, rather than using a cross-validation scheme [6], we put a step forward to directly evaluate on the real future crime data. Specifically, as the cross-validation scheme can remove the fluctuation in the real world data by perform repeated trails on historical data, it fails to evaluate the predictability of the learned model on the *future* crime incidents. Therefore, to maintain the effectiveness of the learned crime prediction model, we define a testing period in the future in which our model is assumed to be valid. Specifically, assuming the training data is built based on the crime data before time t, we now build the testing data using the same method based on the crime data within $[t, t + \Delta]$, where Δ defines the testing period (e.g., one day, one week or one month). Figure 8 shows the example of defining training and testing data over time. In other words, we consider the crime prediction model trained based on data before time t to be valid only during $[t, t + \Delta]$, and retrain a new model at $t + \Delta$ for the next testing period. Therefore, Δ defines the update frequency of our crime prediction model. We will later study the crime prediction performance with different Δ, and empirically select an appropriate updating frequency Δ.

Figure 8 Example of building training and testing datasets over time

5 CrimeTelescope prototype

In this section, we present the implemented prototype of CrimeTelescope with particular focus on its user interface. As all the system components and algorithms are running on the back-end, users need a simple yet informative interface to visualize the crime prediction results. Therefore, the user interface should be designed to provide a user-friendly visualization of the potential risk of crimes in a city. To achieve this goal, we resort to heat maps to visualize the predicted crime probabilities. As shown in Figure 9, the user interface is built on top of Google Maps. The advantage of using heat maps is that the crime hotspots can be easily identified by their color.

A control panel with multiple options is also provided to users for further exploration. First, on the top of the map, users are able to choose a time period (the current week or one past month) for displaying the corresponding crime hotspot maps. Then, on the left panel, users can choose to visualize the predicted hotspot map, or the historical hotspot map (if the corresponding historical crime data is available). To let users investigate the crime hotspots of different crime types, a selector of crime types is also provided. Finally, we provide two
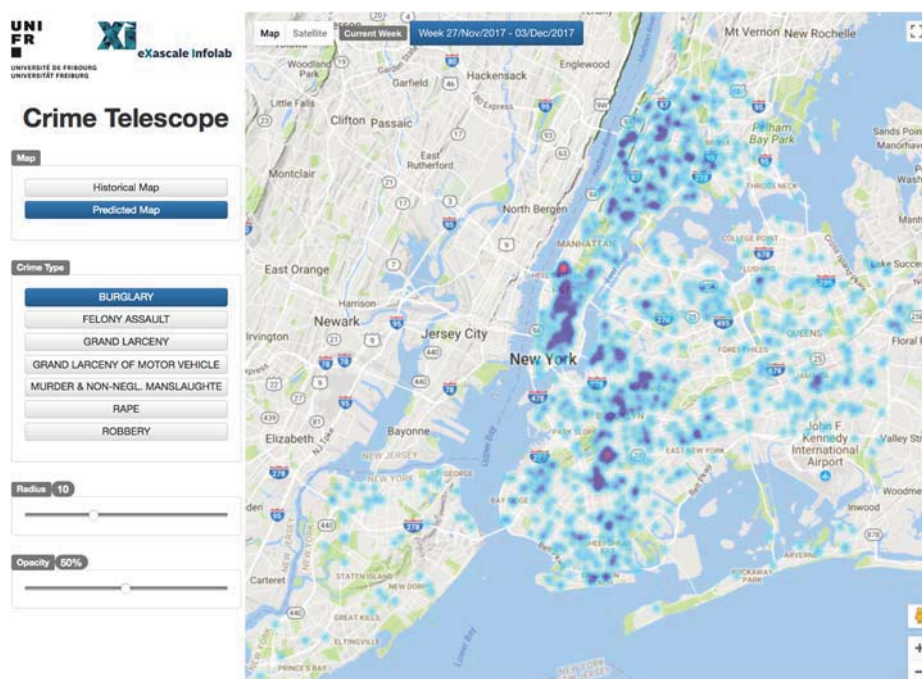

Figure 9 User interface of CrimeTelescope prototype (Accessible online here: http://prediction.heaney.ch/)

options to customize the visual effects of the hotspot map: "Radius", which controls the size of the smoothing area to generate the heat map, and "Opacity", which controls the transparency degree of the heat map layer. The prototype of CrimeTelescope is accessible online[14].

6 Evaluation of crime prediction

In this section, we quantitatively evaluate our crime prediction approach based on real-world data. In the following, we first present our experimental setup, including the dataset, baselines and metrics. We then compare our method with different data fusion schemes. Finally, we discuss the crime prediction results across different crime types and its implications.

6.1 Experimental setup

In order to evaluate our crime prediction approach, we use data collected from Jan. 2014 to Apr. 2015, and perform the crime prediction for each testing period Δ (every day, every week, every month) in the first four months in 2015. Specifically, we use the crime data, city infrastructure data and Tweets data prior to a test period as training data for prediction. More information on the data collection process can be found in Section 4.2.

To validate our approach, we compare it with the following baselines:

– **KDE**. This first baseline is based on a classical crime prediction approach, which is based on historical crime data only [11]. The estimate from KDE is the only feature, i.e., F_l^c. In our experiments, we run KDE based on crime data from the previous month or the previous year, which are denoted as (last month) and (last year), respectively.
– **POI**. This baseline uses only the features extracted from urban infrastructure data F_l^p for crime prediction. Here, we focus on the top two levels of POI categories, denoted as (lv1) and (lv2), respectively (see Section 4.2).
– **Tweet**. This baseline uses only the features extracted from Tweets F_l^t for crime prediction. Here, we perform the feature extraction based on Tweets observed in 30, 10 or 1 day prior to a testing period, which are denoted as (daily rolling 30), (daily rolling 10) and (daily rolling 1), respectively.
– **KDE+POI**. This baseline merges features extracted from historical crime data and urban infrastructure data for crime prediction, where the estimate from KDE and the distribution of POI categories are aggregated as features, i.e., $[F_l^c, F_l^p]$.
– **KDE+Tweet**. This baseline merges the features extracted from historical crime data and Tweets for crime prediction. The estimate from KDE and the features extracted from Tweets are aggregated as features, i.e., $[F_l^c, F_l^t]$.
– **KDE+Tweet+POI**. Our proposed approach merges the features extracted from the three data sources, i.e., $F_l = [F_l^c, F_l^t, F_l^p]$. The configuration for individual data sources will be discussed in the evaluation results.

As we formulate the crime prediction problem as a classification problem, we conduct experiments with popular classification algorithms, including Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), neural network (NN), Decision Tree

---

[14] http://prediction.heaney.ch/

(DT), and Random Forests (RF). The hyperparameters for each classifier are empirically tuned using cross-validation to achieve best results. Specifically, we tune the following parameters for individual classifiers. For feature preprocessing, we tune each classifier with/without feature standardization. For training classifiers, we use MATLAB default hyperparameter optimization (if available) for individual classifiers: 1) LR: regularization term strength; 2) NB: data distributions including kernel smoothing density estimation with different smoothing window sizes, normal distribution, multinomial distribution, multivariate multinomial distribution; 3) SVM: box constraint penalizing margin-violating observations, and kernel scale parameter; 4) NN: hidden layer size; 5) DT: minimum number of leaf node observations; 6) RF: number of trees. To evaluate classification performance, we use three common metrics, i.e., accuracy, F1 score and Area Under Curve (AUC) [22]. Higher values of these metrics imply better performance. Note that as the focus of this paper is on data fusion rather than classification algorithms, we report only the best results from individual classifiers.

6.2 Comparison between different data fusion schemes

In this section, we compare the overall crime prediction results of different data fusion schemes. The testing period Δ is set to one week, and we compute the average results of all test periods in the first four months in 2015. As mentioned previously, we built one model for each crime type, and perform predictions using the corresponding model. The reported results are then the weighted average of all crime types, where the weights are the ratio of individual crime types in the testing data. Figure 10 shows the results of comparing different data fusion schemes with different metrics and classification algorithms.

First, we observe that models based on features extracted from single data source show unsatisfied results over all classification methods, compared to data fusion methods. When comparing results using individual data sources (KDE, POI or Tweet), we find that historical crime data (i.e., KDE) results in the best performance in general, which indicates that the historical crime data is indeed the primary data source for crime prediction [6, 11, 16]. We now discuss the different settings for individual data sources. When comparing the two configurations for KDE, we find that the KDE (last year) achieves better performance than KDE (last month). This is probably due to the fact that the historical crime data of one month are sparse, and are not able to accurately reflect the crime hotspots. Comparing the two levels of POI categories, we find that POI (lv2) consistently achieves better results than POI (lv1), as finer-grained POI categories can more subtly characterize urban infrastructure. Comparing the three configurations when adding Tweet data, we find that Tweet (daily rolling 1) generally achieves the best results, followed by (daily rolling 10) and by (daily rolling 30). The results show that the more recent Tweet data have higher predictive power. Hence, we select KDE (last year), POI (lv2) and Tweet (daily rolling 1) in the following experiments.

Second, we find that when fusing KDE with urban infrastructure data (i.e., POI data) or with Tweet data, we always observe improvements on all three metrics over all classification methods. Furthermore, our approach considering all three data sources consistently outperforms all other baselines. Our results show that the fusion of various urban and social media data is able to improve the overall crime prediction performance. Particularly, compared to the classical KDE method that leverages only the historical crime data, our approach based on data fusion is able to effectively improve the prediction performance. For example, we observe an increase of accuracy by $3.4 - 5.2\%$ across different classification methods.
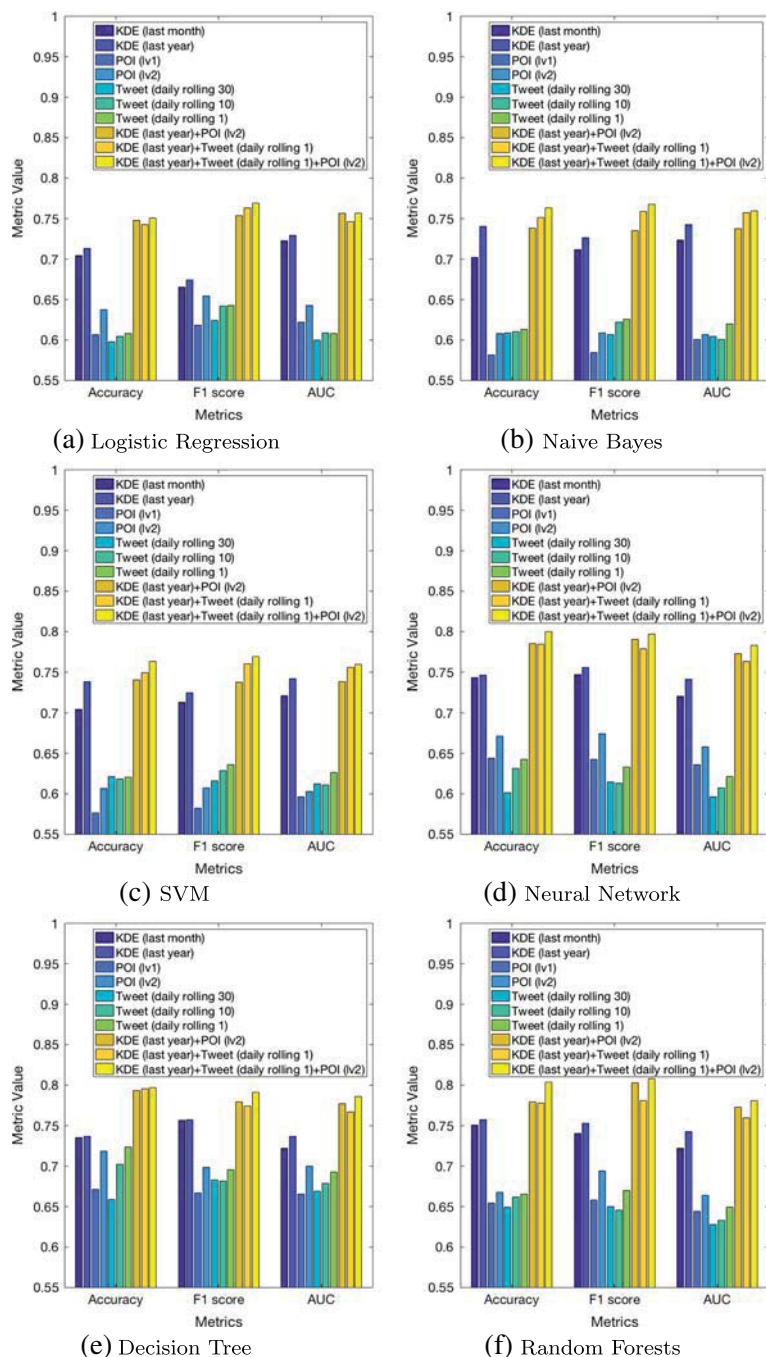
(a) Logistic Regression

(b) Naive Bayes

(c) SVM

(d) Neural Network

(e) Decision Tree

(f) Random Forests

Figure 10 Prediction performance comparison between different data fusion schemes

18

6.2.1 Implication of different features on the learned model

To further understand the implication of our crime prediction model, we focus on the logistic regression method. Specifically, one key advantage of using a regression model is that the weight assigned to each feature can be interpreted as the contribution of that feature to the prediction result. In our case of crime prediction, such a weight corresponds to the extent that a feature will increase the probability of a location being a crime scene. In order to have a deeper understanding of the crime prediction problem, we investigate such weights in the regression model. Unsurprisingly, we find that the estimates from KDE always have the highest positive weight, which indicates that the historical crime data is indeed the primary data source for crime prediction [6, 11, 16], i.e., future crimes often occur in the vicinity of areas that had a high concentration of crime activities in the past.

Moreover, we investigate the weights assigned to individual POI categories (at the second level) by the learned logistic regression model, and find some interesting results. On one hand, the top two POI categories with the highest positive weights are "Argentinian Restaurant" and "Mexican Restaurant", which implies that the presence (or higher density) of POIs of these two categories will probably lead to a prediction of crime incidents. On the other hand, the top two POI categories with the highest negative weights are "College Auditorium" and "College & University", which implies that the presence (or higher density) of POIs of these two categories will reduce the probability of crime incidents there.

6.3 Performance over time

In this section, we study the crime prediction performance of our approach over time. Specifically, we conduct experiments with different updating frequency Δ, i.e., one day, one week, one month. In other words, we retrain a new model every day, every week or every month in order to maintain its effectiveness. Similar to the previous experiment, the reported results are the weighted average of all crime types, where the weights are the ratio of individual crime types in the testing data. To better visualize the results, we need to choose a unified granularity over time. As ground truth crime data has a strong fluctuation on a daily granularity, we report the results for each week in the first four months in 2015. More precisely, for the case of Δ being one day, we update the model everyday and evaluate it on the next day, and then report the average result for each week. For the case of Δ being one month, we update the model every month and evaluate it individually on the following weeks. We report the results using Random Forests, which shows the best performance in the previous experiments. We also note that we obtain similar results using other classifiers.

Figure 11a, b and c show the accuracy, F1-score and AUC over time, respectively. Since we evaluate our model on the real future crime data, we observe that the weekly results show a certain fluctuation. For a clear comparison with different Δ, we also plot the average performance over time in Figure 11d.

When comparing the results with different update frequency Δ, we observe that frequent updates can efficiently maintain the effectiveness of the learned crime prediction model. Specifically, comparing the cases of $\Delta = one\ day$ and $\Delta = one\ month$, we observe similar performance at the beginning of a month (e.g., 1st week). Then, their performance difference increases over time (e.g., 2nd to 4th week), caused by the fact that the model of $\Delta = one\ month$ become less and less effective over time. Finally, after updating the model

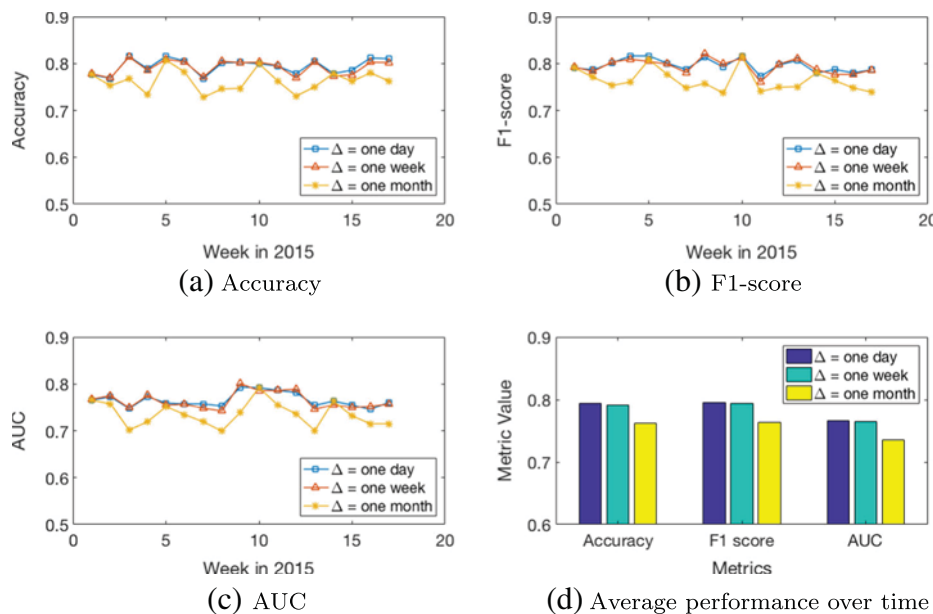(a) Accuracy   (b) F1-score

(c) AUC   (d) Average performance over time

Figure 11 Prediction performance over time for different testing periods Δ

at the end of the first month for $\Delta = one\ month$, its performance increases to the similar level as for $\Delta = one\ day$ (e.g., 5th week).

In addition, we find similar performance for Δ=*one day* and Δ=*one week*, which means that the learned model can keep its effectiveness within the next week. On the other hand, an update of the model involves all feature extraction process, which is computational expensive. Therefore, we empirically set Δ=*one week* in other experiments to maintain the model effectiveness.

6.4 Comparison between crime types

Different crime types have been shown to exhibit different hotspot patterns [16]. In this section, we compare the results obtained from different crime types using our approach. As shown in Figure 12, we observe different predictabilities across the crime types, which are relatively consistent across different classification algorithms. In order to understand such differences, we further investigate the crime hotspot maps of two types of crimes, i.e., "Rape" with the highest accuracy and "Grand Larceny of Motor Vehicle" with a lower accuracy. As illustrated in Figure 13, we observe that the hotspots of crime type "Rape" are fewer and more concentrated than those of "Grand Larceny of Motor Vehicle". The more concentrated the hotspots are, the higher prediction accuracy we can obtain, as highly concentrated (i.e., less uncertainty over space) crime activities are usually easier to predict.

Moreover, such results also motivate us to re-consider the influencing factors for individual crime types. Specifically, the same factor extracted from urban and social media data may have a different impact on the crime prediction of different crime types. We plan to explore this direction and design type-specific crime prediction approaches in future work.
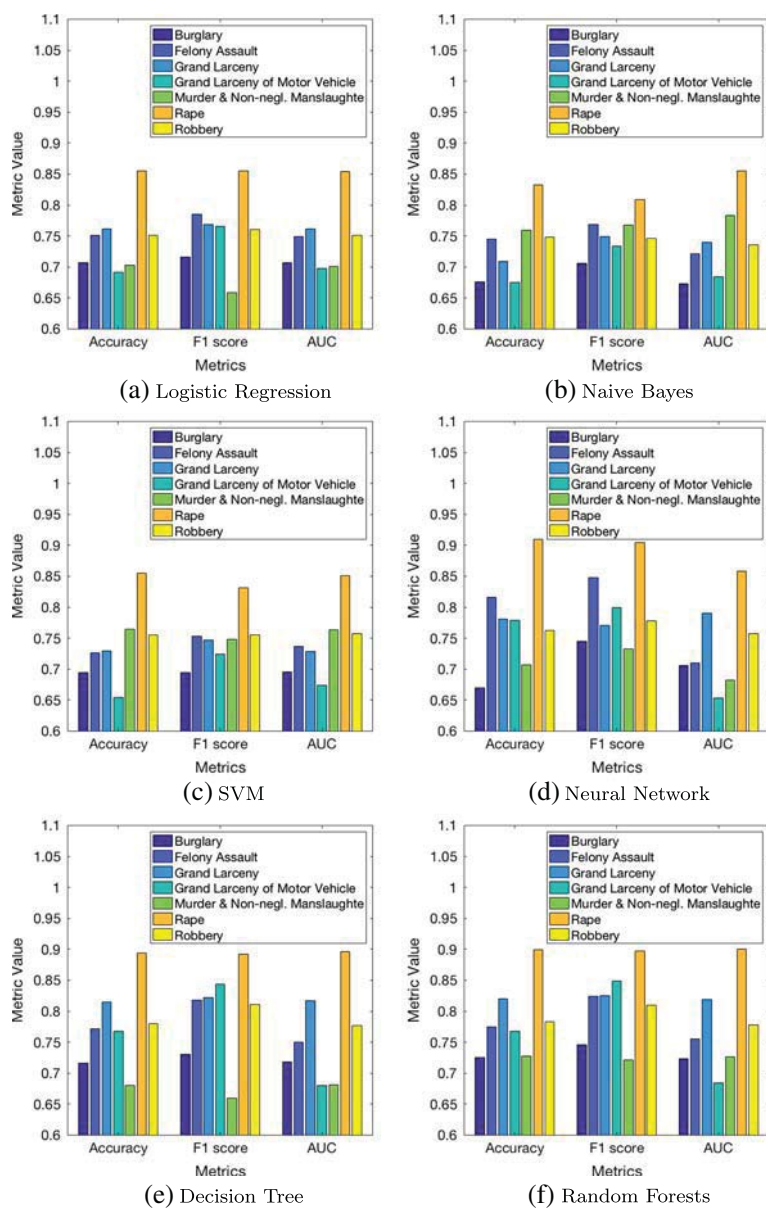
20

Figure 12 Prediction performance comparison between different crime types

7 System usability scale survey

One of the important design goals of CrimeTelescope is to provide a simple yet informative user interface to visualize crime hotspots. Therefore, in order to evaluate our platform from a user experience perspective, we use the System Usability Scale (SUS) survey [10], which is a popular usability evaluation method that has been used to assess various online
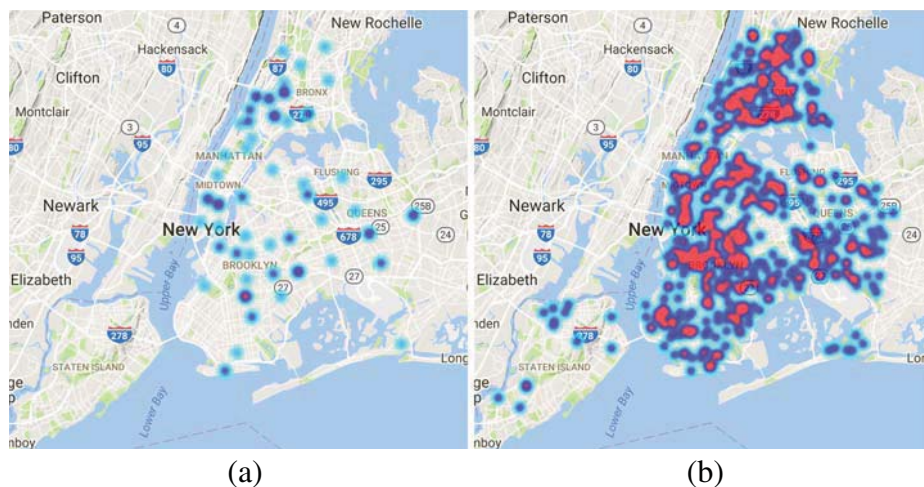
Figure 13 Crime hotspot map for two crime types: **a** Rape; **b** Grand Larceny of Motor Vehicle

systems, such as clinical decision support systems [34], search engines [14], disaster management systems [43], social network applications [44], etc. This SUS survey contains a questionnaire with ten statements, where participants are supposed to indicate the degree of agreement with each statement according to the Likert Scale [21], i.e., a five-point scale from 1 ("Strongly disagree") to 5 ("Strongly agree"). The ten statements is organized in the way that the odd-numbered items (1, 3, 5, 7 and 9) are positive statements and the even-numbered ones (2, 4, 6, 8 and 10) are negative statements. Table 4 shows the statements of the SUS survey.

After the completion of the survey, the results are analyzed as follows. First, a score contribution from 0 to 4 is determined for each statement. Specifically, the score contribution of

Table 4 System Usability Scale (SUS) survey with the resulted scores

SUS Statements	Score
S1: I think that I would like to use this system frequently.	2.77
S2: I found the system unnecessarily complex.	3.58
S3: I thought the system was easy to use.	3.64
S4: I think that I would need the support of a technical person to be able to use this system.	3.58
S5: I found the various functions in this system were well integrated.	3.15
S6: I thought there was too much inconsistency in this system.	3.46
S7: I would imagine that most people would learn to use this system very quickly.	3.62
S8: I found the system very cumbersome to use.	3.46
S9: I felt very confident using the system.	3.30
S10: I need to learn a lot of things before I could get going with this system.	3.58
Learnability dimension (S4 and S7)	3.60
Usability dimension (other 8 statements)	3.37
Overall SUS score	85.35

Note that the scores for S1-S10, learnability and usability range from 0 to 4, while the overall SUS score is between 0 to 100. For all the cases, higher scores imply better user experience

a positive item is the scale position minus 1, while the score contribution of a negative item is 5 minus the scale position. Subsequently, a higher score for a positive items implies more agreement on its statement, while a higher score for a negative item implies less agreement on its statement. Afterwards, the SUS survey calculates a single score indicating the overall usability of the system. It is computed by multiplying the sum of the score contributions by 2.5 (the overall SUS score is thus between 0 to 100). Higher scores imply better user experience. Furthermore, by conducting factor analysis on the SUS statements, Lewis et al. [20] further define two finer-grained dimensions, i.e., learnability and usability, where the learnability dimension contains the statements 4 and 10, and the usability dimension contains the statements 1, 2, 3, 5, 6, 7, 8, and 9.

We conducted our survey of CrimeTelescope platform using Google Forms[15]. The participants are recruited by emails, websites and social networks. A brief description of our platform is provided to participants. After putting the survey online for three weeks, 26 participants were recruited in total, of which 7 were female. Most of the participants were between 20 and 30 (11 participants) and $30 - 40$ (9 participants) years old. The participants have diverse professions, including researchers, software/hardware engineers, students, professors, account managers, etc.

The survey results is shown in Table 4 (including scores for individual statements, learnability, usability, and the overall SUS score). First, the overall SUS score is 85.35, corresponding to an "excellent" SUS rating according the mapping of SUS scores (from 0 to 100) to adjective ratings (i.e, worst imaginable, awful, poor, OK, good, excellent, best imaginable) [2]. Moreover, our platform achieves high scores for both learnability and usability, of 3.60 and 3.37, respectively. Note that the learnability and usability scores range from 0 to 4, and higher scores imply better user experience.

To better understand the usability of our platform, we further investigate the results of individual statements. We find that S2, S3, S4, S7 and S10 have relatively high scores, while S1 has a low score. On one hand, these high scores indicate that the user interface of CrimeTelescope is well designed and easy to use without further technical support nor specific preliminary knowledge. On the other hand, we find that participants' agreement on "using the system frequently" is relatively low. By analyzing the comments they left in the survey, we understand that some users are willing to use the system only when necessary, such as when traveling or moving to a new place. Those are very reasonable use cases for citizens; In addition, we believe that CrimeTelescope can serve as a powerful decision support system for urban authorities and police departments.

8 Conclusions and future work

In this paper, we described CrimeTelescope, an online crime prediction and visualization platform based on urban and social media data fusion. Specifically, CrimeTelescope can continuously collect heterogeneous urban and social media data (i.e., historical crime data, urban infrastructure data and Tweet data), and extract key features from them. By aggregating the extracted features, CrimeTelescope learns a classification model for crime prediction. Finally, it offers a simple yet compelling Web-based user interface for visualizing crime hotspots on an interactive map. To evaluate CrimeTelescope, we first focused on crime prediction accuracy based on real-world data, and then assessed its system usability

---

[15]https://docs.google.com/forms

based on a SUS survey. The results show that CrimeTelescope can achieve accurate crime prediction via data fusion. Particularly, compared to classical approaches based only on historical crime data, CrimeTelescope is able to improve the prediction accuracy by up to 5.2%. The results of the SUS survey show that CrimeTelescope provides a user-friendly interface for crime hotspot visualization.

In the future, we plan to extend CrimeTelescope to incorporate other major cities around the world. Moreover, we intend to include other types of urban and social media data—such as demographics and social event data—to further improve the prediction accuracy. Finally, we also want to explore the temporal dynamics of criminal activities to further provide fine-grained crime hotspot prediction.

References

1. Arulanandam, R., Savarimuthu, B.T.R., Purvis, M.A.: Extracting crime information from online newspaper articles. In: Proceedings of the second australasian Web conference, pp. 31–38. Australian Computer Society, Inc., Sydney (2014)
2. Bangor, A., Kortum, P., Miller, J.: Determining what individual sus scores mean: adding an adjective rating scale. J. Usability Stud. **4**(3), 114–123 (2009)
3. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., Sebastopol (2009)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
5. Block, R.L., Block, C.R.: Space, place and crime: hot spot areas and hot places of liquor-related crime. Crime and place **4**(2), 145–184 (1995)
6. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., Pentland, A.: Once upon a crime: towards crime prediction from demographics and mobile data. In: Proceedings of the 16th international conference on multimodal interaction, pp. 427–434. ACM, New York (2014)
7. Braga, A.A., Hureau, D.M., Papachristos, A.V.: An ex post facto evaluation framework for place-based police interventions. Eval. Rev. **35**(6), 592–626 (2011)
8. Brantingham, P.J., Brantingham, P.L.: Environmental criminology. Sage Publications, Beverly Hills (1981)
9. Brantingham, P.L., Brantingham, P.J.: A theoretical model of crime hot spot generation. Studies on Crime & Crime Prevention (1999)
10. Brooke, J.: Sus-a quick and dirty usability scale. Usability Evaluation in Industry **189**, 194 (1996)
11. Chainey, S., Tompson, L., Uhlig, S.: The utility of hotspot mapping for predicting spatial patterns of crime. Secur. J. **21**(1), 4–28 (2008)
12. Ehrlich, I.: On the relation between education and crime. In: Education, income, and human behavior, pp. 313–338. NBER, Massachusetts (1975)
13. Eysenck, H.J.: Crime and personality (Psychology Revivals). Routledge, Abingdon (2013)
14. Ferré, S., Hermann, A.: Semantic search: reconciling expressive querying and exploratory search. In: International semantic Web conference, pp. 177–192. Springer, Berlin (2011)
15. Garland, D.: Governmentality'and the problem of crime: foucault, criminology, sociology. Theor. Criminol. **1**(2), 173–214 (1997)
16. Gerber, M.S.: Predicting crime using twitter and kernel density estimation. Decis. Support. Syst. **61**, 115–125 (2014)
17. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224n Project Report, Stanford **1.2009**, 12 (2009)
18. Huang, Y.Y., Li, C.T., Jeng, S.K.: Mining location-based social networks for criminal activity prediction. In: Wireless and optical communication conference (WOCC), 2015 24th, pp. 185–189. IEEE, Piscataway (2015)
19. Kitchin, R.: The real-time city? big data and smart urbanism. GeoJournal **79**(1), 1–14 (2014)

20. Lewis, J.R., Sauro, J.: The factor structure of the system usability scale . In: Human centered design, pp. 94–103. Springer, Berlin (2009)
21. Likert, R.: A technique for the measurement of attitudes. Arch. Psychol. **22**(140), 1–55 (1932)
22. Ling, C.X., Huang, J., Zhang, H.: Auc: a better measure than accuracy in comparing learning algorithms. In: Conference of the canadian society for computational studies of intelligence, pp. 329–341. Springer, Berlin (2003)
23. Lynch, A.K., Rasmussen, D.W.: Measuring the impact of crime on house prices. Appl. Econ. **33**(15), 1981–1989 (2001)
24. Mukherjee, S. et al.: Ethnicity and crime. Trends Iss. Crime Crim. Justice **117**, 1 (1999)
25. Newton, A., Felson, M.: Crime patterns in time and space: the dynamics of crime opportunities in urban areas. Crime Sci. **4**(1), 1–5 (2015)
26. Raphael, S., Winter-Ebmer, R.: Identifying the effect of unemployment on crime. J. Law Econ. **44**(1), 259–283 (2001)
27. Ratcliffe, J.H., Taniguchi, T., Groff, E.R., Wood, J.D.: The philadelphia foot patrol experiment: a randomized controlled trial of police patrol effectiveness in violent crime hotspots. Criminology **49**(3), 795–831 (2011)
28. Scott, D.W.: Multivariate density estimation: theory, practice, and visualization. Wiley, Hoboken (2015)
29. Shuyo, N.: Language detection library for java. http://code.google.com/p/language-detection/ (2010)
30. Signorini, A., Segre, A.M., Polgreen, P.M.: The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. PloS one **6**(5), e19,467 (2011)
31. Sun, Y., Wong, A.K., Kamel, M.S.: Classification of imbalanced data: a review. Int. J. Pattern Recognit. Artif. Intell. **23**(04), 687–719 (2009)
32. Taylor, B., Koper, C.S., Woods, D.J.: A randomized controlled trial of different policing strategies at hot spots of violent crime. J. Exp. Criminol. **7**(2), 149–181 (2011)
33. Toole, J.L., Eagle, N., Plotkin, J.B.: Spatiotemporal correlations in criminal offense records. ACM Trans. Intell. Syst. Technol. **2**(4), 38 (2011)
34. Trafton, J., Martins, S., Michel, M., Lewis, E., Wang, D., Combs, A., Scates, N., Tu, S., Goldstein, M.K.: Evaluation of the acceptability and usability of a decision support system to encourage safe and effective use of opioid therapy for chronic, noncancer pain by primary care providers. Pain Med. **11**(4), 575–585 (2010)
35. Traunmueller, M., Quattrone, G., Capra, L.: Mining mobile phone data to investigate urban crime theories at scale. In: Social informatics, pp. 396–411. Berlin, Springer (2014)
36. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Election forecasts with twitter: How 140 characters reflect the political landscape. Social Sci. Comput. Rev. **29**(4), 402–418 (2011)
37. Vivacqua, A.S., Borges, M.R.: Taking advantage of collective knowledge in emergency response systems. J. Netw. Comput. Appl. **35**(1), 189–198 (2012)
38. Wang, H., Kifer, D., Graif, C., Li, Z.: Crime rate inference with big data. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 635–644. ACM, New York (2016)
39. Wang, T., Rudin, C., Wagner, D., Sevieri, R.: Learning to detect patterns of crime. In: Machine learning and knowledge discovery in databases, pp. 515–530. Springer, Berlin (2013)
40. Wang, T., Rudin, C., Wagner, D., Sevieri, R.: Finding patterns with a rotten core: data mining for crime series with cores. Big Data **3**(1), 3–21 (2015)
41. Wang, X., Gerber, M.S., Brown, D.E.: Automatic crime prediction using events extracted from twitter posts. In: International conference on social computing, behavioral-cultural modeling, and prediction, pp. 231–238. Springer, Berlin (2012)
42. Yang, D., Zhang, D., Yu, Z., Wang, Z.: A sentiment-enhanced personalized location recommendation system. In: Proceedings of the 24th ACM conference on hypertext and social media, pp. 119–128 (2013)
43. Yang, D., Zhang, D., Frank, K., Robertson, P., Jennings, E., Roddy, M., Lichtenstern, M.: Providing real-time assistance in disaster relief by leveraging crowdsourcing power. Pers. Ubiquit. Comput. **18**(8), 2025–2034 (2014)
44. Yang, D., Zhang, D., Chen, L., Qu, B.: Nationtelescope: monitoring and visualizing large-scale collective behavior in lbsns. J. Netw. Comput. Appl. **55**, 170–180 (2015)
45. Yang, D., Zhang, D., Qu, B.: Participatory cultural mapping based on collective behavior data in location-based social networks. ACM Trans. Intell. Syst. Technol. (TIST) **7**(3), 30 (2016)
46. Yuan, J., Zheng, Y., Xie, X.: Discovering regions of different functions in a city using human mobility and pois. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 186–194. ACM, New York (2012)
47. Zhong, Y., Yuan, N.J., Zhong, W., Zhang, F., Xie, X.: You are where you go: Inferring demographic attributes from location check-ins. In: Proceedings of the 8th ACM international conference on Web search and data mining, pp. 295–304. ACM, New York (2015)