Aghababaei and Makrehchi *Hum. Cent. Comput. Inf. Sci.* (2017) 7:3
DOI 10.1186/s13673-016-0084-z

Human-centric Computing and Information Sciences

RESEARCH Open Access

Activity-based Twitter sampling for content-based and user-centric prediction models

Somayyeh Aghababaei* and Masoud Makrehchi

*Correspondence: somayyeh.aghababaei@uoit.ca
Department of Electrical, Computer and Software Engineering, University of Ontario Institute of Technology (UOIT), 2000 Simcoe St N, Oshawa, ON L1H 7K4, Canada

Abstract

Increasingly more applications rely on crowd-sourced data from social media. Some of these applications are concerned with real-time data streams, while others are more focused on acquiring temporal footprints from historical data. Nevertheless, determining the subset of "credible" users is crucial. While the majority of sampling approaches focus on individual static networks, dynamic user activity over time is usually not considered, which may result in activity gaps in the collected data. Models based on noisy and missing data can significantly degrade in performance. In this study, we demonstrate how to sample Twitter users in order to produce more credible data for temporal prediction models. We present an activity-based sampling approach where users are selected based on their historical activities in Twitter. The predictability of the collected content from activity-based and random sampling is compared in a content-based and user-centric temporal model. The results indicate the importance of an activity-oriented sampling method for the acquisition of more credible content for temporal models.

Keywords: Twitter sampling, Temporal prediction models, Historical timelines, User activity, Activity-based sampling

Background

Twitter's public and open nature provides great opportunities for its users to actively participate in sharing their opinions and produce high quality content that is reflective of their tendencies and preferences in their day-to-day life [1]. This vast amount of publicly available user-generated content is applied to many applications ranging from tracking human social behavior [2–4], detecting events of interest [5–7], to smart business [8] where domain knowledge is collected through social media. These studies are either concerned with pulling Twitter and aggregating tweets as bulk or tracking historical tweets over time in order to find meaningful patterns for targeted events. The main challenge of the former studies is the limitation of the Twitter API in accessing only 1% of all existing tweets. However, despite this limitation, the latter studies are concerned with retrieving historical timelines of users.

To tackle the above issues of retrieving more tweets beyond the 1% threshold and obtaining historical timelines, topic-based sampling and REST API are both shown to

be more effective [9, 10]. In topic-based sampling [11], a set of specific keywords or hashtags are applied to collect tweets through the search API. A very substantial problem with this group of sampling is that it is limited to the studies around the content of shared topics, which is not scalable to many applications. In contrast to topic-based models, the REST API can be a user-based scenario, which provides access to user history. In the case of the REST API, a set of Twitter users are needed in order to retrieve historical tweets. However, the issue of selecting a credible subset of users still remains. Nevertheless, many network-based sampling approaches were studied, which focus on sampling a subset of users from their networks [12] or sampling users based on their popularity [13]. The drawback behind the network-based sampling is that, a set of users are sampled from a static network while ignoring the availability of their posts over time. In fact, there is no guarantee that sampled users are active on a daily basis, which is necessary for temporal models.

In this study, we sample Twitter wherein, we propose an activity-based sampling method to retrieve a selection of users for the REST API. In activity-based sampling, we leverage user profiles to extract their historical activities. The most active users are assumed as "credible" users for employing in a temporal prediction model. We address two main characteristics in our sampling model: (a) obtaining the most active users, (b) avoiding missing content or activity gaps over time. The term active users does not refer to celebrities, news agencies, or major companies whose corporate accounts in social media are normally managed by a group of employees.

We gathered two samples of Twitter users using our proposed sampling approach and random users. The random users refer to users who post in real-time, which are collected using streaming API. Since streaming API is widely used approach in many topical and user-based models [14–16], it is important to assess the effectiveness of the activity-based sampling proposed in this study compared with random sampling. The selected users from both approaches are employed in the REST API to collect their historical tweets. We compare the content of users, selected from both sampling approaches in different aspects, including statistical properties and predictability in temporal models.

We employ the collected historical content in two temporal prediction models; user-centric and content-based—they both aim to discover conclusions from user-generated content. In the user-centric model, the content of a set of selected users is aggregated based on user timelines, to extract meaningful patterns [17], while in the content-based approach, the content of all individuals are combined together with respect to the event of interest [18]. Both of the aforementioned approaches are considered to be temporal models, which suffer from the challenge of retrieving tweets over time. In a temporal model where content is tracked to detect a set of patterns, the availability of tweets over time significantly effect the model performance. This approach suffers from activity gaps or missing data. Therefore, we can evaluate the effectiveness of our proposed sampling compared with the random approach in providing more credible content while mitigating the effect of missing content. Overall, the data gathered from the activity-based and random sampling are compared in three main aspects:

(a) *Timelines* Do the samples provide enough data for the consideration period of time?
(b) *User activity* How the samples covered the period of interest. Do we observe missing posts over time?
(c) *Content credibility* How retrieved content is effective for the temporal user-centric and content-based models?

The rest of the paper is organized as follows: "Related works" section presents a background of existing Twitter sampling approaches. "Sampling approaches" section describes the proposed method for Twitter sampling. We present two prediction models to evaluate the data collected from the proposed sampling approaches in "Temporal classification model" section. In "Dataset" section, we discuss the characteristics of data collected by two sampling approaches along with the results obtained for the prediction. The conclusion and the limitations of the current study along with our future works are presented in "Experimental results" section.

Related works

With the increasing number of Twitter users, the size of tweets have become overwhelming and Twitter sampling, the selection of subset of tweets or users, is particularly relevant. Many sampling techniques were studied ranging from topical [11, 19] to user-based approaches [12]. The first set of techniques is topic-based sampling, where specific keywords or hashtags are applied to collect tweets through Twitter API [6, 20]. As an example, Kumar and Geethakumari [19] used different keywords to collect tweets related to natural calamities and political event for the purpose of detecting misinformation in Twitter. This group of sampling limits the study around the content of shared topics, which are not scalable to many applications. The second group focuses on sampling a subset of users from their networks [21]. The drawback behind the latter approach is that, the availability of user posts over time is not considered. In fact, there is no guarantee that sampled users are active on a daily basis, which is necessary for temporal models where content (content-based) [18] or user timelines (user-centric) are aggregated considering their timestamps [17].

The most common sampling approach is random sampling using streaming API, which allows retrieving 1% of real-time data with some specific parameters. There have been many empirical studies dealing with the evaluation of the data sampled from random sampling with other approaches, including random versus firehouse [22]. This study discusses the situations in which random sampling has less coverage compared to firehouse. However, when there is more specific parameters such as keywords, random sampling can provide "enough" data as firehouse. In another study [11] streaming API is compared with expert sampling. The expert users are the users with high number of followers. In this study, content of expert users were compared with random users in terms of trustworthy of their content. It is revealed that expert content contains more divers and popular topics and includes less spam, which has application in many topical extraction models such as breaking news detection. Therefore, we can conclude from

previous studies and the recent ones [23] that expert sampling is rich in content and is more valuable for content-based models such as topical models. In fact, Twitter streaming preserves the statistics of the sample size as the whole representative sample, but for content-based models which can benefit from the context, expert sampling is more superior. Hence, streaming API is highly depend on the type of coverage and the targeted problem.

Although, many empirical studies evaluated the effectiveness of expert sampling in many dimensions such as trustworthy, diversity of discussion topics, statistic representative of samples, or sentiment. However, there are many challenges in utilizing content of experts, whose corporate accounts in social media are normally managed by a group of employees, compared to random users. In many applications, ranging from content-based [24] to user-centric [17], opinions of crowds collectively provide predictive signals for prediction models. In fact, by filtering experts we ignore the valuable content coming from crowd and we neglected the vast amount of information contributed by the citizens.

A vast amount of studies prefers network sampling rather than selection of experts based on popularity. In network sampling, a subset of users are chosen from the entire network of collected users for perfect sampling. Different techniques have been applied in recent years, of which Random Walk and Breadth-First Search (BFS) [25] are well-known. However, the major problem with the mentioned techniques is that, these techniques are biased toward high degree nodes similar to expert sampling. A solution to this problem is the traditional Monte Carlo Markov Chain (MCMC), which was proposed by White et al. [12]. They applied a technique based on MCMC and Coupling From The Past (CFTP) to have better convergence in sampling. These methods ignore the activity of users over time, whereas in temporal models, the presence of users over time is mostly needed.

In temporal models such as detecting target events [26], discovering spatio-temporal topics [27, 28], or tracking user behavior over time [29], user activity or content shared over time is tracked to extract meaningful signals. Therefore, activity gaps or missing opinions can significantly degrade the performance of both content-based and user-centric models. Although many sampling approaches were presented to select a subset of users and content in static mode, there is a significant need for a sampling approach to address the temporal aspect of data. In this study, we investigate how to retrieve users to decrease the activity gaps. We also investigate how much retrieved content from sampled users are effective for a temporal prediction model. In fact, we leverage user profiles to estimate their activities in the past and to determine the most active users as opposed to expert users.

Sampling approaches

The objective of this study is to present a sampling approach to collect the best representative users for the REST API. In contrast to often used Streaming API, the REST API can be a user-based approach with less limitation to access Twitter data. Given a set of users, the REST API provides access to historical timelines, with the limitation of at most 3200 recent tweets for a single user. The main challenge is how to sample Twitter users to avoid the absence of data in historical tweets. Nevertheless, absent data could

be inevitable, users do not necessarily share posts on a daily basis. However, as far as possible, to avoid missing opinions in historical tweets, we address some characteristics for the selection of users. In this method, the interest is to find a set of the most active users while showing no bias toward individuals with a high or low number of tweets. We collect users selected by two different sampling strategies; a random approach using the streaming API and an activity-based sampling which is based on the historical activity of a user. The use of the network-based sampling is not considered in this study due to the nature of the targeted problem. In this study, we are looking for independent opinions, while the network sampling (users and their networks) is biased toward the same opinions.

Random sampling

As discussed earlier, random sampling is the most common approach to access data streams. In order to obtain random users, we gathered 1% of tweets using streaming API. The historical timeline of the randomly selected Twitter users are retrieved using the REST API.

Activity-based sampling

In this method, the interest is to find a set of active users while being unbiased to individuals with very high or low numbers of tweets. In our sampling approach, two factors are considered: the period of time a user is active and its daily number of tweets. Since these specifications are not available, we retrieve them from user profiles. For each tweet, user profile of its author is retrieved, which includes some fields such as: status_count and created_at. For each user, two main specifications are calculated as follows:

1. The number of days a users is active (*days*). In order to understand for how many days a user is active, we calculate the number of days the user's profile was generated till the current time. A longer period of activity is a primary criteria for the selection. As we track the content of users over time, users who recently became members are ignored.
2. The average number of tweets per day (*tweets_day*): As this parameter is irretrievable, we leverage the total number of tweets for the user and the number of days a user is active

$$tweets\_day = total\_tweets/days \tag{1}$$

where we assume a user has uniform activity behavior. A user is considered active if it has a high number of active days (*days*) and a high number of tweets per day (*tweets_day*). The active users are classified by using the numbr of followers to filter out accounts belonging to celebrities, news agencies or major companies.

Temporal classification model

In this section, we introduce two temporal prediction models: content-based and user-centric. In both models, historical content shared by Twitter users are applied for prediction. In the content-based approach, the prediction model employs data, aggregated across all users, as bulk. However, in the user-centric model, the content of the selected users is aggregated based on user timelines for a selected task. The task is crime trend

prediction by leveraging Twitter data. Historical tweets have shown to be successful in predicting the directions of crime rates [17, 18]. The problem of trend prediction is converted to a binary classification problem where the objective is to detect the directions of the targeted trends (in our case, crime trends). Previous studies [30] shown that the classification approach is effective in predicting the occurrence of different crimes. However, in this study, we address the changes of crime rates. In fact, the prediction problem is transformed into a supervised classification task that predicts whether crime rates increase or decrease for the prospective timeframe.

Prior to classification, a set of N training documents of the form $\{(x_1, l_1), (x_2, l_2), \ldots, (x_N, l_N)\}$ are generated in which x_i is the feature vector of the i-th document and l_i is its assigned label. For the purpose of creating documents $X = \bigcup_{i=1}^{N} x_i$, two different approaches are applied; concatenation of the content for the content-based approach and the aggregation of user opinions for the user-centric model. The generated documents are then associated with a set of labels. The labels are inferred from the knowledge obtained from the targeted problem (here crime index), which is the directions of rates in the prospective timeframe. Although the labels are inferred in a same manner for both models (content-based and user-centric), they have different strategy of generating documents. Figure 1 shows the framework of the data generation for both models as well as the timeline. After Twitter users were sampled, they are fed to the REST API to retrieve historical timelines of the selected users. The collected data along with the crime rate directions are employed in the content-based and user-centric predictions, which are discussed in the following subsections. Table 1 presents the notations being used in this section.

Content-based model

As discussed earlier, in the content-based model, documents are generated based on timestamps of tweets posted by all users without consideration of filtering any users. In fact, this model captures collective patterns from the crowd rather than a selected group

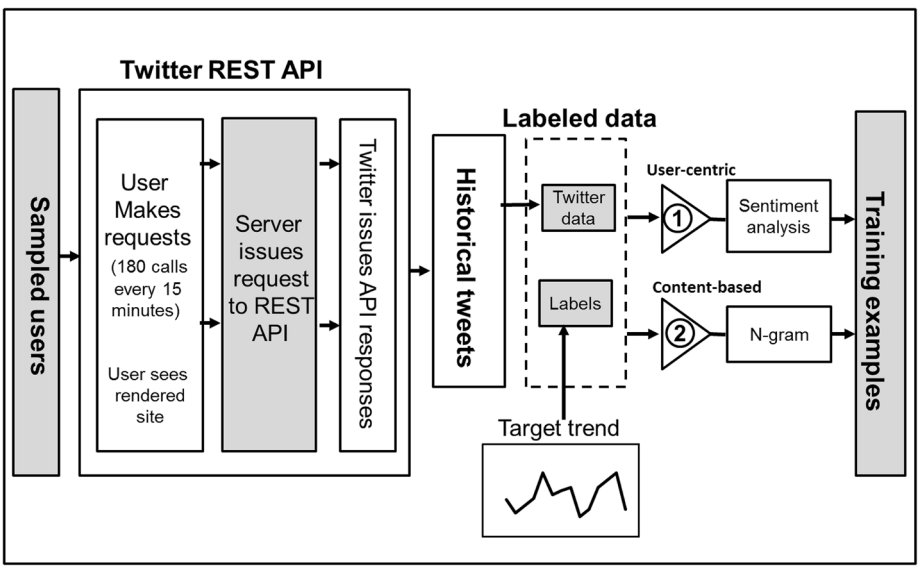

Fig. 1 The framework of the temporal model for the user-centric (see *1*) and content-based (see *2*) prediction models

Table 1 Notations of variables and their corresponding description

Notations	Descriptions
N	Total number of documents
M	Total number of users, $1 < m < M$
V	Global vocabulary
w	Word in vocabulary
u_i	ith user out of m
s_u	Sentiment score belongs to user (u)
q	Aggregation window
y_i	Crime rate at time $t(i)$
Δr	Lag between a document and a target trend
p_i	A post tweeted at time $t(i)$
d_i	A document sampled at time $t(i)$
$\chi^{\langle c \rangle}$	Document term matrix of size $N * \|V\|$ sparse matrix
$\chi^{\langle u \rangle}$	Document sentiment matrix of size $N * M$ sparse matrix

of users. All observed users are considered as crowd, as opposed to the user-centric approach. In order to generate training examples a set of temporal document are generated. Let $d_i = \{p_1^{\langle i \rangle}, p_2^{\langle i \rangle}, \ldots, p_m^{\langle i \rangle}\}$ denotes a document, which consist of a set of posts shared at time $t(i)$. Let $D = \{d_1, d_2, \ldots, d_n\}$ be a set of temporal documents or in general temporal data, which is defined as a state in time. The state is represented by vector of features $d_i = (f_1, f_2, \ldots, f_{|V|})$, where V is the global vocabulary. Since each state d_i is sampled at time $t(i)$, then $D = \bigcup_{i=1}^{n} d_i$ is the result of n consecutive sampling. One important pre-processing task in time-series data, is smoothing to increase the predictability and to reduce the noise and outliers. Hypothetically, temporal data which is a high-dimensional time-series data can be also smoothed. In our model, each state is represented by a document and a naive smoothing is a rolling averaging algorithm over the temporal documents;

$$x_i^{\langle c \rangle} = \frac{1}{q} \sum_{j=1}^{q} d_{j-q+1}^{\langle c \rangle}, X^{\langle c \rangle} = \bigcup_{i=1}^{n} x_i^{\langle c \rangle}, \quad q = [1, n] \tag{2}$$

$$X^{\langle c \rangle} = \bigcup_{i=1}^{n} x_i^{\langle c \rangle} = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,|V|} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,|V|} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N,1} & w_{N,2} & \cdots & w_{N,|V|} \end{pmatrix} \tag{3}$$

where q is the size of aggregation window and d_i is an example in $t(i)$ or in our case the day i, which is represented by a single document. All relevant tweets are aggregated into a signal document without targeted filtering. As a result X is an $n \times |V|$ document-term matrix (Eq. 3) where V is the global vocabulary. The vocabulary V is simply a set of all distinct words appeared in all collected, relevant tweets. Although, no keyword search is conducted, a blind filtering including stopword reduction and low-frequent term reduction is applied to the vocabulary. As a result, x_i is defined as the average of a set of documents from j to day $j - q + 1$, retrospectively.

Several preprocessing tasks such as low frequent term and stopword removal may be applied to x_i. In the content-based approach, documents are represented with terms as features, which are referred to N-gram model without filtering any specific keywords. One might speculate that we must collect keywords to emphasize on offensive language implying a rough context. Nevertheless, content is a rich data which contains valuable hidden variables including activities, topic of discussions, public interests, and sentiments, which might not be necessarily carried by offensive language.

User-centric model

In the second model, instead of data aggregation across all users, documents are generated from the individual opinions in different time slots [17]. If a user u_1 has a post at time t and user u_2 also posted something at the same time, the content of each is employed as a unique feature or an user-dependent feature rather than combining them together. Let $T_u = \left\{ \left(p_1^{\langle u \rangle}, t_1^{\langle u \rangle}\right), \left(p_2^{\langle u \rangle}, t_2^{\langle u \rangle}\right), \ldots, \left(p_J^{\langle u \rangle}, t_J^{\langle u \rangle}\right) \right\}$ denotes a timelines of a user u, where tuple $\left(p_j^{\langle u \rangle}, t_j^{\langle u \rangle}\right)$ represents user u's post j along with its timestamps: $t_1^{\langle u \rangle} < t_2^{\langle u \rangle} < \cdots < t_J^{\langle u \rangle}$. Post $p_j^{\langle u \rangle} = \{w_1^{\langle u \rangle}, w_2^{\langle i \rangle}, \ldots, w_{K_j}^{\langle u \rangle}\}$, is comprised of tokens $w_{jk}^{\langle u \rangle}$. $V = \bigcup_{u,j,k} w_{jk}^{\langle u \rangle}$ is a global vocabulary, $k \in [1, K_j^{\langle u \rangle}]$. In order to aggregate tweets based on user timelines, we assume an aggregation window in which user timelines are concatenated as follows:

$$
\begin{aligned}
d_m^{\langle u \rangle} &= \tfrac{1}{q} \sum_{j=1}^{q} p_{j-q+1}^{\langle u \rangle}, \quad q = [1, n] \\
x_i^{\langle u \rangle} &= \left(d_{i,1}^{\langle u \rangle}, d_{i,2}^{\langle u \rangle}, \ldots, d_{i,M}^{\langle u \rangle}\right) \\
X\langle u \rangle &= \bigcup_{i=1}^{n} x_i^{\langle u \rangle}
\end{aligned}
\tag{4}
$$

where q is the size of aggregation window, M is the total number of users, d_m is a timelines of a user after aggregation, and x_i is a document consist of a series of user timelines. Therefore, features vectors are represented as follows:

$$
X^{\langle u \rangle} = \bigcup_{i=1}^{n} x_i^{\langle u \rangle} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,M} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N,1} & u_{N,2} & \cdots & s_{N,M} \end{pmatrix} \tag{5}
$$

where $s_{i,m}$ is the sentiment of the user m, which belongs to document i. Since the idea of this model is considering a sample of users representative of the sentiments of ll users, we used LIWC [31] to detect the sentiment. We extract the positiveness and negativeness scores. Each user is defined by the normalized mentioned scores.

Generating labels

Let $Y = \{y_1, y_2, \ldots, y_n\}$ be the target time series whose future values are to be predicted. The time series Y is sampled in time steps $t(i)$, $1 \leq i \leq n$. To convert regression-based prediction into classification, the continuous signal Y has to be mapped into a categorical

set which is called the set of labels. There are several techniques to infer labels from a continuous variable such as quantization or direction of changes in rates. Due to the nature of the research, we adopt trend analysis of the continuous rates for labeling:

$$l_i = sgn(y_{i+\Delta r} - y_i), \quad if \begin{cases} \Delta r > 0: \ lag \\ \Delta r \leq 0: \ lead \end{cases}, \quad L = \bigcup_{i=1}^{n} l_i \tag{6}$$

where Δr is the lead or lag from current state (x_i) and target label, l_i is the label at $t = i$ and L is the sequence of labels in n consecutive time steps. After inferring labels, a set of annotated examples is generated by associating high dimensional temporal data to one dimensional target labels inferred from time series of interest,

$$\forall x_i \in X, x_i \rightarrow l_i, D = \left\{(x_1, l_1), (x_2, l_2), \ldots, (x_{N-\Delta r}, \ l_{N-\Delta r})\right\}. \tag{7}$$

The objective of the proposed method is to predict whether the trend of interest increases or decreases for the perspective time-frame. Therefore, a set of training data (D) is given to a binary classifier as follow:

$$D = \{(x_i, l_i) | x_i \in R, l_i \in \{-1, 1\}\}, \quad 1 \leq i \leq N - \Delta r \tag{8}$$

where in our target problem (crime trend prediction) z_i is learning documents and the label (l_i) is derived from the changes in crime index when comparing the current index (i) with the index of ($i + \Delta r$), where Δr is the time interval such as one day or 1 week:

$$l_i = \begin{cases} 1 & if \ rate(i) < rate(i + \Delta r) \\ -1 & \text{otherwise} \end{cases} \tag{9}$$

where rate (i) and rate ($i + \Delta r$) are crime index at i and $i + \Delta r$ according to our historical data.

Dataset

In this study we tackle crime prediction as a case study. The idea is how to predict crime rate changes from the tweets posted earlier. We collected Twitter data and crime rates from Chicago, Illinois between January 2014 and October 2015. Chicago has been targeted due to its importance as the third populous city in U.S as well as being among top three cities, which attracted the highest number of visitors during 2012.[1] It has been also ranked as the first in the number of murders, second in robbery, and third in the number of property crimes based on FBI report during 2013.[2]

Crime data

The criminal records were extracted from Chicago Data Portal.[3] This Data Portal is a rich resource providing all reported incidents on a daily basis, which are retrieved from Chicago Police Department system. Information of frequent crimes that have been reported between January 2014 and October 2015 were collected. Each record contains

[1] http://en.wikipedia.org/wiki/Chicago.

[2] S. Department of Justice, FBI: http://www.fbi.gov.

[3] City of Chicago Data Portal: https://data.cityofchicago.org.

its timestamps, exact location, and crime type. The dates refer to the time of primary investigation, and crime type derived based on the FBI classification system. Figure 2 presents the crime rate time series (aggregated rates of all different crime types). A major decrease of overall crime rates observed during the entire period of time which is started in US in 1990s [32]. The significant changes in the number of incidents are coincided with important holidays such as New Year's day and Christmas. However, they might be the result of missing data. A major decrease of overall crime rates is observed during the entire period of time which is started in US in 1990s [32].

Twitter data

In order to retrieve historical Twitter data, two sets of Twitter users were collected using the random and activity-based sampling as discussed in "Sampling approaches" section. Historical timelines of the selected users were retrieved and restricted to the same timeframe—between January 1, 2014 and October 1, 2015.

Figure 3 presents the number of selected active users over 30 days for the activity-based sampling. The figure indicates two different trends: "Unseen" stands for the number of active users who are selected each day (d), and " Seen" represents the number of users labeled as active but already selected for d. As can be observed, the number of new active users who are not detected decreases over time. Due to the increase of repeated users, the process of collecting active users was terminated after almost 1 month ($d = 30$). We applied the REST API to retrieve their historical timelines of the selected users. Historical timelines of the users were restricted to the same timeframe of crime rates—between January 2014 and October 2015.

Experimental results

In this section, we evaluate how much the proposed sampling approaches can minimize the lack of data and deliver more informative content. The historical timelines of the selected users from two different approaches; activity-based and random sampling are retrieved using the REST API. We evaluate the feasibility of our sampling approach

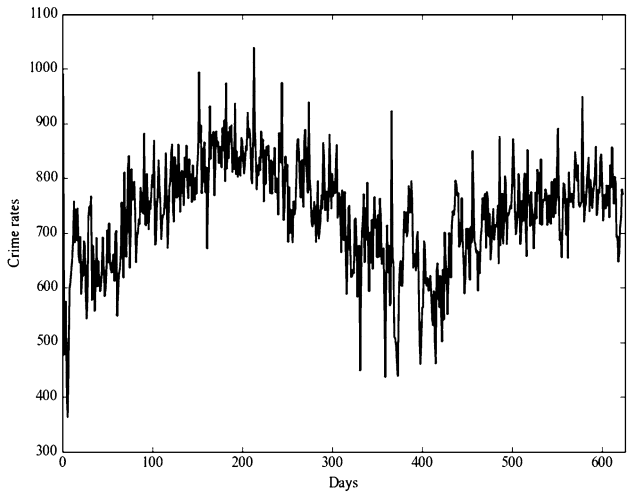

Fig. 2 Daily number of crime rates

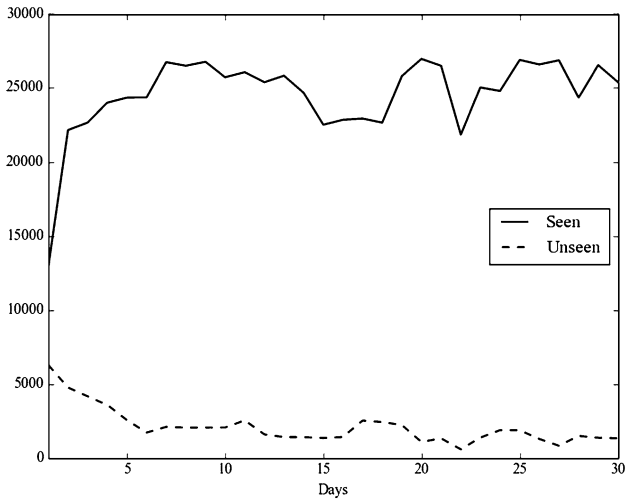

Fig. 3 Daily number of users over 30 days

compared with the random sampling in retrieving historical tweets. We begin with comparing statistical characteristics of data collected from both approaches. The intention is to understand how well data are distributed over time for both sampling approaches. We then evaluate the credibility of the content in the proposed prediction temporal models.

Comparing timelines

We compare the number of posts (see Fig. 4) and users (see Fig. 5) observed on a daily basis from both datasets. The historical tweets obtained from the active and random users are mapped between our consideration period of time using their timestamps, we did not go back more than 600 days because of the low number of activities. As a result, we reached tweets during January 2014–October 2015. Figure 4 presents that the daily number of tweets from the active users are higher than tweets of the random users. This

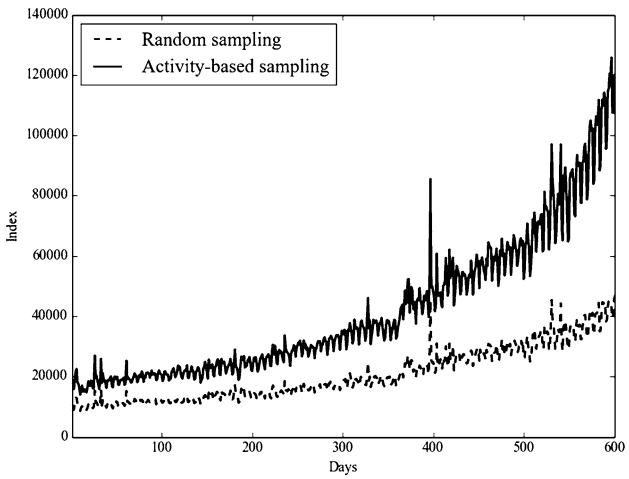

Fig. 4 Daily number of tweets captured from activity-based and random datasets

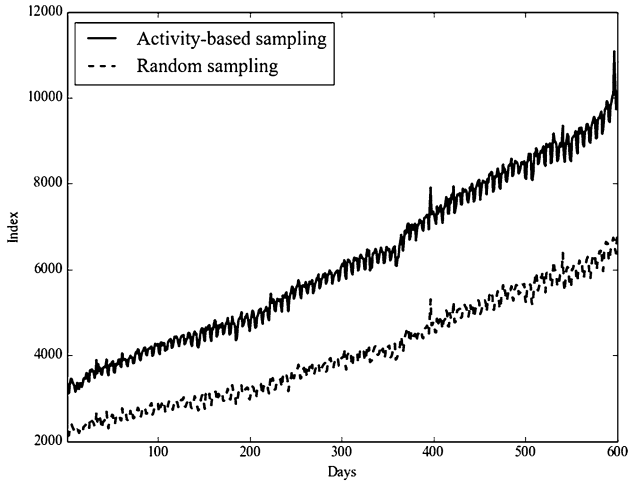

Fig. 5 Daily number of active users captured from activity-based and random datasets

can be an asset for content-based models where availability of content is crucial for the performance of a temporal topical model. For topical models where a set of parameters such as: keywords or hashtags is retrieved from the collected content, this way a sampling approach with more coverage might be able to extract more data over time. Figure 5 shows the daily number of unique users, defined as those who post at least once per day. From Fig. 5 we can observe that the daily number of active users obtained from the activity-based is higher than the number of users from the random sampling. For each day, a higher number of users were active for the activity-based sampling compared to the random. In user-centric models, the number of available active users plays an important role in providing interesting patterns for targeted problems [33, 34]. The statistics of historical tweets and users show that the activity-based sampling compared to the random has better coverage in terms of number of tweets and users. In fact, One of the key question of this paper was how to efficiently capture historical tweets which then is applied for content-based and user-centric temporal models. Content-based models are challenged with the number of tweets available on daily basis and in user-centric approaches, the number of available active users plays an important role.

In more details we are also interested to know how many tweets each individual has posted. In general, the REST API has a limitation of providing only 3200 (or slightly higher) number of tweets of a specific user. However, if the targeted user did not post more than 3200 tweets, we can retrieve entire timelines of the selected user. Figures 6 and 7 show the distribution of overall posts between users for the random and activity-based sampling respectively. The Figures capture very interesting pattern that most of the users (5350) in the activity-based approach are active and have more than 3000 number of tweets. Surprisingly, many users in the random sampling were not active during the selected period of time. They were mostly active only during the time data were collected and has no contributions during the past. More than 3000 users from the random users had <100 tweets during the past. The selected users in the random sampling do not have long time contribution in posting tweets. They were mostly active during data collection, therefore, the numbers of historical tweets were not significant.

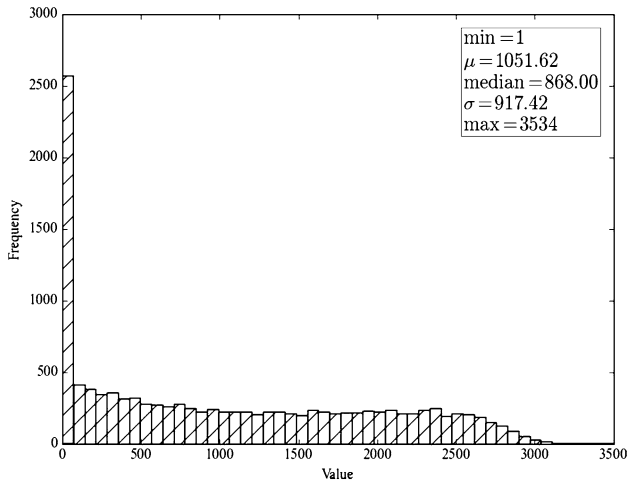

Fig. 6 Distribution of overall posts between users for random sampling

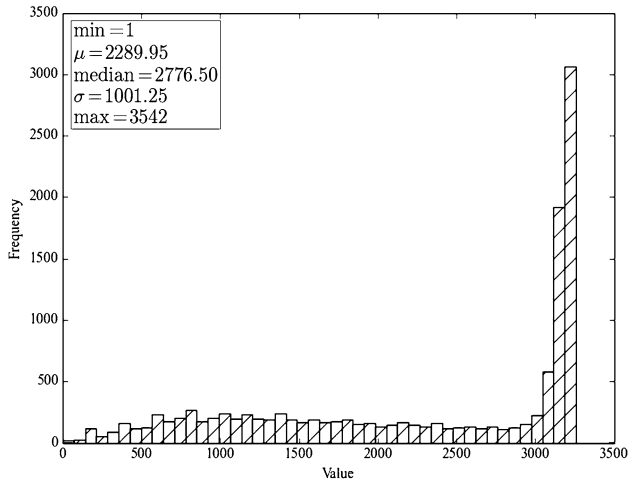

Fig. 7 Distribution of overall posts between users for activity-based sampling

Comparing activity gaps

We also investigate the presence of user activity over time, which is the key element in user-centric approaches. Models directly working with user streams are prone to vast amounts of missing opinions. The absent data are related to the errors occurring during data collection or simply users are not active during a specific period of time. Although activity gaps are inevitable, it is crucial to retrieve the most active users while avoiding activity gaps in their timelines. While random sampling ignores the activity of the selected users during the past, the activity-based sampling selects users based on their historical timelines. Figures 8 and 9 show the daily activity of the 100 most active users during 632 days for the activity-based and random sampling respectively. In this figures, The indexes of the users (y axis) were plotted against the period of time (x axis). The vertical black bar indicates a user has at least one tweet in that specific days where the white space shows the absence of the user. In fact, the figures indicate the activity of each user

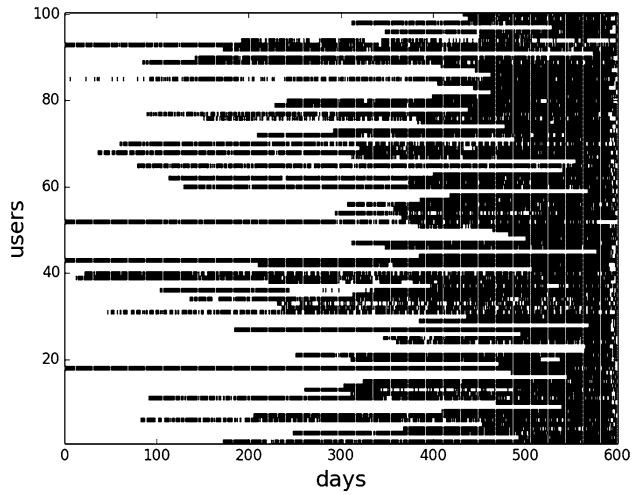

Fig. 8 Rastergram of daily activity by 100 most active users for activity-based sampling

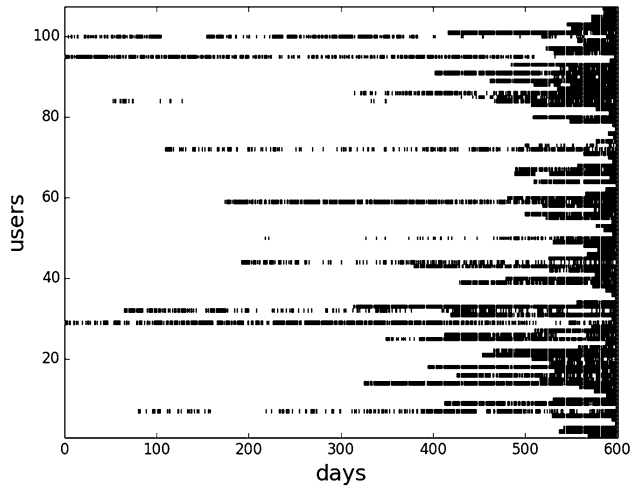

Fig. 9 Rastergram of daily activity by 100 most active users for random sampling

over the consideration period of time. Although the top 100 active users, who posted the highest number of tweets, were selected from both approaches; the activity gaps in random sampling (Fig. 9) is inevitable. It can be due to the selection of users based on their activity in the streaming time not based on their history. However, from the Fig. 8, we can observe that the activity-based significantly reduces the absent data, which are more applicable in user-centric approaches.

Comparing credibility

We evaluated the credibility of the datasets in prediction models. The predictability of the content extracted from active users was compared with the content retrieved from random users in two models: the content-based and user-centric approaches. As discussed before ("Temporal classification model" section), both models are temporal classification models with different document generation approaches. The classifier

is linearSVC, which is the implementation of liblinear [35]. LinearSVC is faster compared with LinearSVM, since kernel transforms are not used and it scales better for large datasets in a linear classification problem. The evaluation was processed by calculating the Macro-averaged F1-score and using rolling origin [36] as the common method for training and evaluating the performance of the model for series observations. In this approach, the training set is the first i (80% of the dataset) and it is tested on the $i+1th$ document. In the second iteration, the training set is moved one document forward (the first $i+1$) and it is tested on the $i+2th$ document. This process is continued until all the test data is classified.

The predictability is compared when the content is applied to predict crime trends with different lags. In this regard, document x_i which has been generated at time t_i, is labeled with crime trend l_i with different lags (see Eq. 4). The lag does not stand for a week or a day, it is a window of time in which crime rate directions are captured. As an example, if lag $= 1$ ($\Delta r = 1$), each document is labeled with the direction of the crime trend in a day later. In each different lags, the classifier is fed with the generated training data separately. For instance, Figures 10 and 11 show the crime trend of BATTERY between a period of 14 days and the generated labels (either $+1$ or -1) for lag $= 1$ and lag $= 2$ respectively. In the case of these two different lags, documents are labeled as

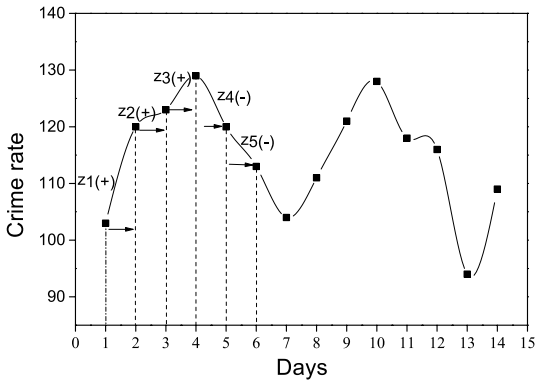

Fig. 10 The labeling approach based on lag = 1

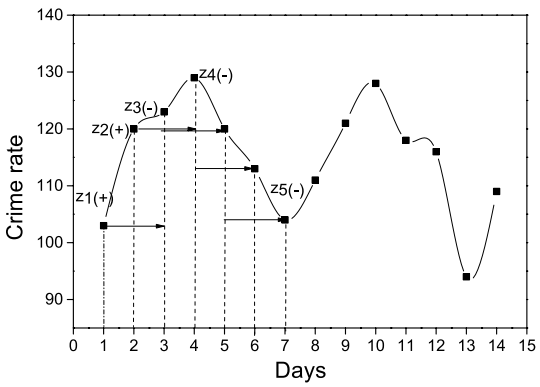

Fig. 11 The labeling approach based on lag = 2

presented in Table 2, where x_1 is a document aggregated at time $t(1)$ and l_1 is its assigned label. The performance of the classifier in lag = 1 and 2 are evaluated separately.

Prediction performance: content-based

As discussed before, the content-based model is based on generating documents from aggregating content as bulk with regards to the tweet timestamps. For pre-processing, we removed stop-word, and low and high frequent words. We also employed chi-squared for feature selection. The documents were examined with binary and tf-idf representations, however, the best results were achieved using binary representation. Table 3 illustrates the F-measure of the prediction for the content-based model where

Table 2 Labeling approach for lag = 1 and lag = 2

Lag = 1	Lag = 2
$z_1 \rightarrow l_1 : sgn\|y_2 - y_1\| = +1$	$z_1 \rightarrow l_1 : sgn\|y_3 - y_1\| = +1$
$z_2 \rightarrow l_2 : sgn\|y_3 - y_2\| = +1$	$z_2 \rightarrow l_2 : sgn\|y_4 - y_2\| = +1$
$z_3 \rightarrow l_3 : sgn\|y_4 - y_3\| = +1$	$z_3 \rightarrow l_3 : sgn\|y_5 - y_3\| = -1$
$z_4 \rightarrow l_4 : sgn\|y_5 - y_4\| = -1$	$z_4 \rightarrow l_4 : sgn\|y_6 - y_4\| = -1$
$z_5 \rightarrow l_5 : sgn\|y_6 - y_5\| = -1$	$z_5 \rightarrow l_5 : sgn\|y_7 - y_5\| = -1$
$z_6 \rightarrow l_6 : sgn\|y_7 - y_6\| = -1$	$z_6 \rightarrow l_6 : sgn\|y_8 - y_6\| = -1$
$z_7 \rightarrow l_7 : sgn\|y_8 - y_7\| = +1$	$z_7 \rightarrow l_7 : sgn\|y_9 - y_7\| = +1$

Table 3 The prediction performance for content-based over 7 lags

	Lag = 1	Lag = 2	Lag = 3	Lag = 4	Lag = 5	Lag = 6	Lag = 7
Activity-based							
Narcotics	*0.51*	*0.54*	0.52	0.53	*0.58*	0.53	*0.67*
Deceptive	*0.65*	0.52	*0.57*	*0.64*	0.65	0.62	0.51
Criminal damage	*0.43*	0.6	*0.7*	0.65	0.6	0.56	*0.54*
Burglary	0.52	*0.58*	*0.56*	*0.56*	0.56	0.54	0.52
Battery	*0.61*	*0.7*	0.62	*0.72*	0.67	0.66	*0.6*
Assault	0.46	0.47	*0.57*	0.52	0.5	*0.54*	*0.56*
Prostitution	0.57	0.59	*0.7*	*0.68*	0.68	0.58	*0.68*
PublicViolation	*0.46*	*0.51*	*0.47*	*0.51*	*0.55*	*0.55*	*0.53*
Robbery	*0.55*	*0.56*	0.47	0.55	*0.52*	*0.52*	*0.56*
Theft	*0.65*	0.55	0.52	*0.6*	*0.62*	0.58	*0.62*
All	*0.77*	*0.74*	*0.7*	*0.86*	*0.76*	*0.7*	*0.73*
Random							
Narcotics	0.5	0.51	*0.57*	*0.55*	0.55	*0.55*	0.65
Deceptive	0.63	*0.56*	0.55	0.55	*0.68*	*0.67*	*0.60*
Criminal damage	0.42	*0.62*	0.69	*0.68*	*0.65*	0.56	0.51
Burglary	*0.53*	0.5	0.52	0.53	*0.57*	0.54	0.52
Battery	0.46	0.67	*0.65*	0.71	*0.7*	*0.7*	0.57
Assault	0.46	*0.54*	0.56	0.54	*0.53*	0.52	0.55
Prostitution	*0.62*	*0.62*	0.67	0.67	0.68	*0.64*	0.54
PublicViolation	0.44	0.47	0.46	0.46	0.47	0.53	0.49
Robbery	0.5	0.54	*0.51*	*0.56*	0.5	0.49	0.44
Theft	0.57	*0.57*	*0.56*	0.53	0.61	0.58	0.55
All	0.5	0.59	0.6	0.59	0.54	0.58	0.61

The italic emphasizes show in which experiments the activity-based or random sampling performed better

both the content of active and random users were employed over 7 lags. The highlighted results indicate which content (activity-based or random) is more credible with respect to a specific lag. As an example, the content of activity-based sampling is more predictive (0.67) compared to the content of random users (0.65) for NARCOTICS when the lag = 7. The results demonstrate that the performance of the activity-based sampling for most of the lags are higher than the random sampling. The activity-based has higher predictability in accumulation of all crime types. The predictability of the activity-based content is 27% higher than the random sampling, which indicates the effectiveness of the activity-based sampling for the content-based model where the objective is to predict the directions of indexes without considering the type of crime. However, in some cases such as BURGLARY and PUBLIC VIOLATION, the difference between the predictability of the two datasets is not considerable. Overall, the results indicate that the proposed activity-based sampling generates more predictive content for ALL and most of the crime types, such as BATTERY, NARCOTICS, and PROSTITUTION, with F-measure up to 0.86.

Prediction performance: user-centric

The same sets of experiments were conducted for the user-centric model in which the intention is to leverage individual timelines for document generation. The documents were presented with normalized sentiment scores as discussed in "Temporal classification model" section. We examined the credibility of documents with different labels. The results were presented in Fig. 12 for all users as well as the top 500 users (Fig. 13). From the results we can observe that in most cases (lags), the content obtained from the activity-based has higher predictability compared with the random sampling. In the best case, "All" crime with $lag = 6$, the activity-based sampling has achieved the F-measure up to 0.85, which is 35% higher than the random sampling. Overall, the content of active users was shown to be more credible for the proposed user-centric model, which can be the result of having fewer activity gaps compared with the random sampling. In fact,

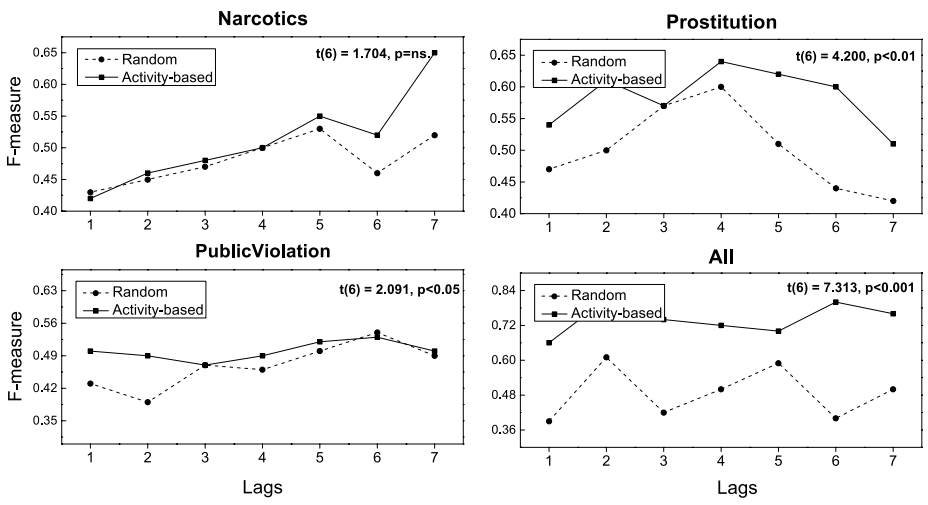

Fig. 12 Predictability for user-centric approach over 7 lags for "All users"

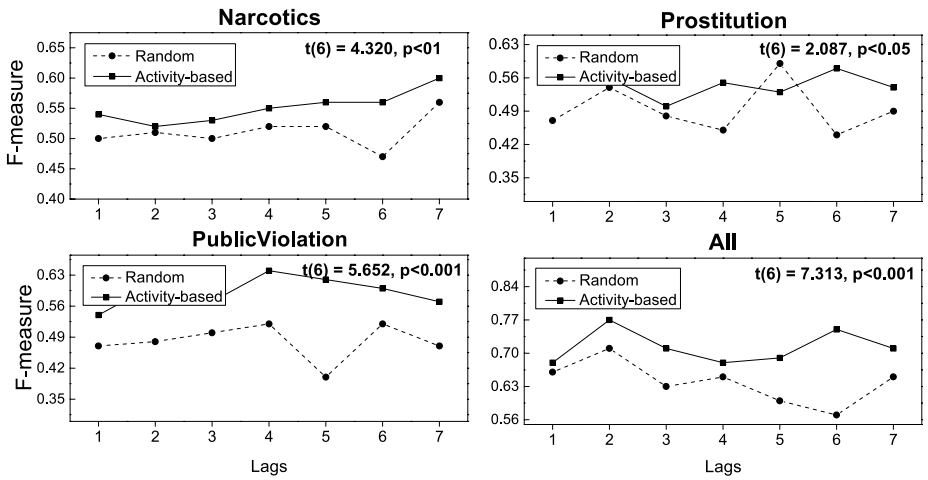

Fig. 13 Predictability for user-centric approach over 7 lags for "Top 500 users"

according to the results, the importance of the activity-based sampling for the user-centric is more significant compared with the content-based model.

Conclusions

Identifying credible sources of content or users are important in many research problems aiming to drive a meaningful conclusion from the source of information. Performance of the prediction models can be degraded from the missing data or the choice of the collected content. This study has argued the importance of the selection of data for the targeted problem. In this work, we focused on sampling Twitter users to retrieve their historical tweets for temporal prediction models. We presented an activity-based approach that leverages user profiles to estimate historical activities in the past for the selection of the most active users as opposed to expert users. In this approach, we selected users based on two factors: the number of days a user is active and the average number of user's tweet per day. Both factors were calculated using user profile elements such as "*created_at*" and "*status_count*". In addition to the activity-based sampling, we also gathered another set of users by random sampling, which is widely used to collect users for the REST API.

The historical timelines of the selected users were also retrieved using the REST API. The timelines of the collected tweets from both groups of users were limited to our period of time consideration. We compared the primary statistical differences between two datasets in terms of historical timelines and user activity. Regarding the number of tweets and users, the activity-based approach has better coverage compared to the random samples. We also compared the overall number of tweets for each user. Most of the users were active (3000 tweets) for the activity-based sampling, and the random users had low activity during the selected period of time. In addition, the activity gap of both sets of users were compared. The results indicate that active users had more contributions in the past, while activity gaps in the random sampling are inevitable. In fact, the activity-based sampling significantly reduces the absent data because users were selected based on their history. Overall, the activity-based approach identifies users

who are more historically active, whereas in the random sampling high activity gaps are observed.

In addition, we also studied the credibility of the content captured from both datasets in the proposed temporal prediction models. We presented two temporal prediction models (user-centric and content-based) to compare the credibility of the content gathered from the selected users. While, in the content-based model, documents are generated based on historical tweets of all collected users, in the user-centric, documents are created based on individual timelines. Both models were applied to predict the directions of crime rates. The prediction models leverage historical tweets to predict crime rate increases or decreases for the prospective timeframe.

The results of the content-based model indicate that the content of active users is more credible in predicting the trend of interest. In the best case, the results is 27% higher when using the content of active users. Overall, in 10 crime types out of 11 the activity-based approach achieved the best results compared to the random sampling. This is due to the fewer activity gaps observed in the collected tweets of the active users compared to the random users. For temporal content-based models such as our proposed model, the availability of content over time plays a crucial role. In the user-centric model, the same performance was observed. The content of active users has higher predictability. In fact, the user-centric model relies on the availability of timelines of users, which is highly affected by the activity gaps. As the results indicated, the performance was significantly higher in some cases (PUBLIC VIOLATION, ALL) compared to the random sampling.

Prior to this research, the properties of different Twitter sampling approaches in different aspects were studied. However, the effectiveness of the collected samples in the targeted domain is less explored. This research has shown the importance of a target-oriented data sampling for prediction models. In addition to the timeline properties and the credibility, we would like to further investigate the quality of the content in terms of discussion topics and sentiments to semantically analyze textual content and the differences in content level. Future work could also address the effectiveness of the proposed sampling approach for other temporal prediction models.

Authors' contributions

SA carried out the acquisition of data, the design and the development of the proposed models. She has been involved in setting up the experiments and interpreting the results. She provided the draft of the manuscript. MM has been involved in the design and development of the proposed models. He has been involved in evaluating the experiments and revising the manuscript critically for important intellectual content. Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 20 September 2016 Accepted: 15 December 2016
Published online: 25 January 2017

References

1. Marwick AE et al (2011) I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media Soc 13(1):114–133
2. Weng J, Lee BS (2011) Event detection in twitter. ICWSM 11:401–408
3. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8
4. Teraoka T (2012) Organization and exploration of heterogeneous personal data collected in daily life. Hum Centric Comput Inf Sci 2(1):1
5. Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on world wide web. ACM, New York, pp 851–860

6. Achrekar H, Gandhe A, Lazarus R, Yu SH, Liu B (2012) Twitter improves seasonal influenza prediction. In: HEALTHINF, pp 61–70
7. Hale S, Gaffney D, Graham M (2012) Where in the world are you? Geolocation and language identification in twitter. In: Proceedings of ICWSMG12, pp 518–521
8. Sato A, Huang R, Yen NY (2015) Design of fusion technique-based mining engine for smart business. Hum Centric Comput Inf Sci 5(1):1
9. Bošnjak M, Oliveira E, Martins J, Mendes Rodrigues E, Sarmento L (2012) Twitterecho: a distributed focused crawler to support open research with twitter data. International World Wide Web Conference Committee (IW3C2), Lyon
10. Zhao WX, Jiang J, He J, Song Y, Achananuparp P, Lim EP, Li X (2011) Topical keyphrase extraction from twitter. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, vol 1. Association for computational linguistics, pp 379–388
11. Ghosh S, Zafar MB, Bhattacharya P, Sharma N, Ganguly N, Gummadi K (2013) On sampling the wisdom of crowds: random vs. expert sampling of the twitter stream. In: Proceedings of the 22nd ACM international conference on information and knowledge management. ACM, New York, pp 1739–1744
12. White K, Li G, Japkowicz N (2012) Sampling online social networks using coupling from the past. In: IEEE 12th international conference on data mining workshops (ICDMW), 2012. IEEE, pp 266–272
13. Ghosh S, Sharma N, Benevenuto F, Ganguly N, Gummadi K (2012) Cognos: crowdsourcing search for topic experts in microblogs. In: Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 575–590
14. Joseph K, Landwehr PM, Carley KM (2014) Two 1% s dont make a whole: comparing simultaneous samples from twitters streaming API. In: International conference on social computing, behavioral-cultural modeling, and prediction. Springer, Berlin, pp 75–83
15. Phuvipadawat S, Murata T (2010) Breaking news detection and tracking in twitter. In: IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT), vol 3, 2010. IEEE. pp 120–123
16. Sofean M, Smith M (2012) A real-time architecture for detection of diseases using social networks: design, implementation and evaluation. In: Proceedings of the 23rd ACM conference on hypertext and social media. ACM, New York, pp 309–310
17. Chepurna I, Aghababaei S, Makrehchi M (2015) How to predict social trends by mining user sentiments. In: International conference on social computing, behavioral-cultural modeling, and prediction. Springer, Berlin, pp 270–275
18. Aghababaei S, Makrehchi M (2015) Temporal topic inference for trend prediction. In: 2015 IEEE international conference on data mining workshop (ICDMW). IEEE. pp 877–884
19. Kumar KPK, Geethakumari G (2014) Detecting misinformation in online social networks using cognitive psychology. Hum Centric Comput Inf Sci 4(1):14. doi:10.1186/s13673-014-0014-x
20. Yun GW, David M, Park S, Joa CY, Labbe B, Lim J, Lee S, Hyun D (2016) Social media and flu: media twitter accounts as agenda setters. Int J Med Inf 91:67–73
21. Leskovec J, Faloutsos C (2006) Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 631–636
22. Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehose. arXiv preprint arXiv:13065204
23. Zafar MB, Bhattacharya P, Ganguly N, Gummadi KP, Ghosh S (2015) Sampling content from online social networks: comparing random vs. expert sampling of the twitter stream. ACM Trans Web 9(3):12
24. Gaurav M, Srivastava A, Kumar A, Miller S (2013) Leveraging candidate popularity on twitter to predict election outcome. In: Proceedings of the 7th workshop on social network mining and analysis. ACM, New York, p 7
25. Kurant M, Markopoulou A, Thiran P (2010) On the bias of BFS (breadth first search). In: 22nd international teletraffic congress (ITC), 2010. IEEE, pp 1–8
26. Wang D, Hom P, Griffeth R, Sager JK (2015) nothing endures but change: investigating temporal dynamics within a turnover model. In: Academy of management proceedings, academy of management, vol 2015, p 15237
27. Yuan Q, Cong G, Ma Z, Sun A, Thalmann NM (2013) Who, where, when and what: discover spatio-temporal topics for twitter users. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 605–613
28. Li L, Goodchild MF, Xu B (2013) Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. Cartogr Geogr Inf Sci 40(2):61–77
29. Hong L, Doumith AS, Davison BD (2013) Co-factorization machines: modeling user interests and predicting individual decisions in twitter. In: Proceedings of the sixth ACM international conference on web search and data mining. ACM, New York, pp 557–566
30. Gerber MS (2014) Predicting crime using twitter and kernel density estimation. Decis Supp Syst 61:115–125
31. Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: Liwc 2001, vol 71. Lawrence Erlbaum Associates, Mahway
32. Mishra S (2014) Crime drop of the 1990s. The encyclopedia of criminology and criminal justice. Blackwell. doi:10.1002/9781118517383.wbeccj410
33. Pennacchiotti M, Popescu AM (2011) Democrats, republicans and starbucks afficionados: user classification in twitter. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 430–438
34. Lampos V, Preotiuc-Pietro D, Cohn T (2013) A user-centric model of voting intention from social media, vol 1. In: ACL. pp 993–1003
35. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: a library for large linear classification. J Mach Learn Res 9:1871–1874
36. Zivot E, Wang J (2003) Rolling analysis of time series. In: Modeling financial time series with S-Plus®. Springer, Berlin, pp 299–346