Two Sides of a Coin: Separating Personal Communication and Public Dissemination Accounts in Twitter

Conference Paper · May 2014

DOI: 10.1007/978-3-319-06608-0_14

6 authors, including:

Conrad Tucker
Pennsylvania State University
96 PUBLICATIONS **493** CITATIONS

SEE PROFILE

Two Sides of a Coin: Separating Personal Communication and Public Dissemination Accounts in Twitter

Peifeng Yin[1], Nilam Ram[2], Wang-Chien Lee[1], Conrad Tucker[3], Shashank Khandelwal[4], and Marcel Salathé[4]

[1] Department of Computer Science & Engineering, Pennsylvania State University
[2] Human Development and Psychology, Pennsylvania State University
[3] School of Engineering Design Technology, Pennsylvania State University
[4] Department of Biology, Pennsylvania State University
{pzy102,wlee}@cse.psu.edu,
{nur5,ctucker4,khandelwal,salathe}@psu.edu

Abstract. There are millions of accounts in Twitter. In this paper, we categorize twitter accounts into two types, namely *Personal Communication Account (PCA)* and *Public Dissemination Account (PDA)*. PCAs are accounts operated by individuals and are used to express that individual's thoughts and feelings. PDAs, on the other hand, refer to accounts owned by non-individuals such as companies, governments, etc. Generally, Tweets in PDA (i) disseminate a specific type of information (e.g., job openings, shopping deals, car accidents) rather than sharing an individual's personal life; and (ii) may be produced by non-human entities (e.g., bots). We aim to develop techniques for identifying PDAs so as to (i) facilitate social scientists to reduce "noise" in their study of human behaviors, and (ii) to index them for potential recommendation to users looking for specific types of information. Through analysis, we find these two types of accounts follow different temporal, spatial and textual patterns. Accordingly we develop probabilistic models based on these features to identify PDAs. We also conduct a series of experiments to evaluate those algorithms for cleaning the Twitter data stream.

1 Introduction

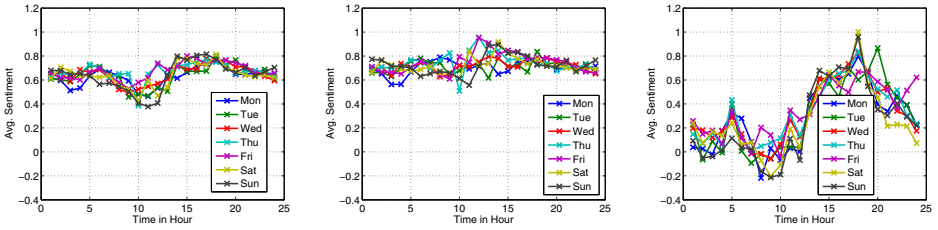
As Twitter[1] has grown, many different kinds of user accounts have emerged. At a general level, we roughly classify them into two categories: (i) Personal Communication Account (PCA) and (ii) Public Dissemination Account (PDA). PCAs are accounts that are usually operated by unique individuals and used for interpersonal communication (e.g., to share personal experiences and opinions). PDAs, in contrast, are typically linked to and operated by a company[2], a web

[1] http://twitter.com
[2] https://twitter.com/#!/citi

V.S. Tseng et al. (Eds.): PAKDD 2014, Part I, LNAI 8443, pp. 163–175, 2014.

site[3] or a program[4] and used to disseminate specific news and information (e.g., locations of car accidents, shopping deals, crimes).

Existence of PDA may cause problems when attempting to study human behavior using Twitter data. Recently, a Twitter-based analysis in *Science* suggested that changes in overall tweet sentiment over time (hours of the day) may be interpreted as evidence of biologically-based diurnal cycles in the mood patterns of humans [7]. However, an underlying assumption is that the tweets were produced by human individuals as part of their natural daily lives. If the data stream is a mixture of PCAs (humans) and PDAs (corporate/entity), the conclusion may be unwarranted. To illustrate it, In Figure 1 we plot the time evolution of average sentiment for 2,787 PCAs and 389 PDAs that were labelled manually. As can be seen, the daily diurnal cycle, i.e., the mood increases in the early morning and decrease later, is easily discernable in the overall data (Figure 1(a)) as similar to the previously published analysis. However, contrary to expectations, we find that the cycles are much less prominent in the PCA (human individual) sub-sample (Figure 1(b)) than in the PDA (corporate/entity) sub-sample (Figure 1(c)). The importance of separating the different account types for reaching accurate conclusions is clear.



(a) Overall (b) PCA (c) PDA

Fig. 1. Sentiment change within a 24 hour time period across different account types

Beyond potentially adding noise to researchers' data streams, PDAs are potentially very useful for various types of data consumers. For instance, an individual looking for jobs may follow PDAs that publish job postings of a particular type (e.g., web development) in a particular geographic area (e.g., New York). Similarly, shoppers may follow PDAs that provide timely notification of nearby sale events and hot deals. In sum, as PDAs' tweets are often focused on a very specific topic and formatted in a uniform manner, they are relatively easy to process and may thus provide rich content for individuals, researchers, and the recommendation engines that support those populations.

The enormous size of the Twitter data stream makes it highly impractical to manually check the account type. In this paper we develop and test a variety

[3] `https://twitter.com/#!/WHERE`
[4] `https://twitter.com/#!/memcrime`

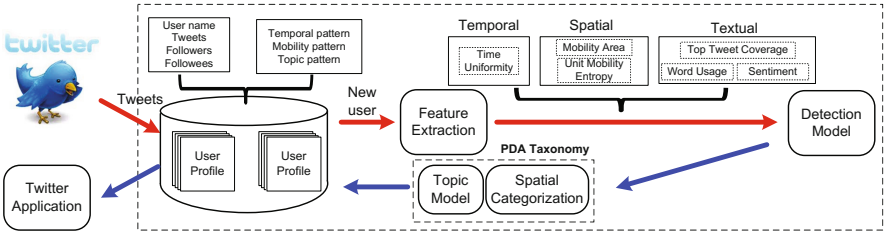


Fig. 2. The framework for PDA detection

of techniques for automatic classification of PDAs and PCAs using multiple temporal, spatial, and textual features of accounts' tweet publishing patterns.

Figure 2 gives an overview of the proposed framework for PDA detection. As shown, tweets are continuously sent to the database. Once a new user arrives, her profile of raw data is checked and different types of features are extracted. Specifically as illustrated in Figure 2, there are temporal, spatial and textual features (details are discussed in Section 2). With extracted features, a classification model is then employed to determine the account type. After a PDA is detected, the system checks its posted tweets to model its topic as well as categorizing its spatial characteristics. Finally, all extracted features and topic models are also saved in the database. Twitter applications, e.g., user recommendation, can then be built upon the knowledge mined in this framework.

The rest of the paper is organized as follows. Section 2 describes the feature extraction. Section 3 provides details of our models. Section 4 reports the evaluation of our model and shows some PDAs found using our model. Section 5 reviews relevant research and finally Section 6 concludes the whole paper.

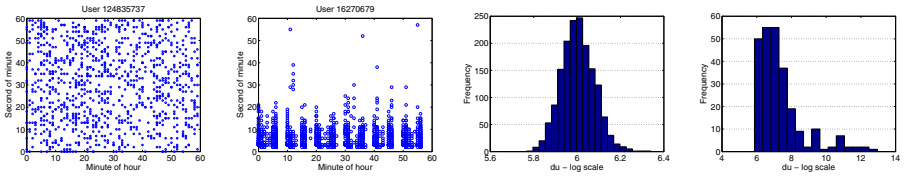
2 Feature Extraction

In this section we discuss the extraction features used to identify PCAs and PDAs. The work makes use of an archive of geo-tagged tweets published between March 1, 2011 and January 18, 2012 [1]. During this time, 39,994,126 geo-tagged tweets (with latitude and longitude attached) were posted by 1,506,937 users. Ground truth classification data were generated by randomly selecting 5,000 accounts that published at least 200 tweets and manually labeling them as PCA, PDA, or unknown. Of the 5,000 randomly selected accounts 2,787 were PCAs, 389 PDAs, 0 spam accounts and 1,824 unknown accounts[5]. These data were then used to extract and analyze the temporal, spatial and textual features of PCAs and PDAs.

[5] In our manual search of the data we did not identify any spam accounts. They may not appear in the geo-tagged tweet stream, perhaps to maintain anonymity, or because Twitter had already detected and blocked the tweet content (in which case we would have labeled them as unknown).

2.1 Temporal Feature

PDAs are, by definition, regularly disseminating useful information. Often this task is facilitated by use of automated computer programs that publish tweets at specific times or at regular intervals [23]. In contrast, PCAs, being human, may be less regimented in their communication of daily live events and feelings. Figure 3(a) and 3(b) show the timing (minutes by seconds) of tweets published by two Twitter accounts. The specific times at which tweets were sent by the user depicted in the Figure 3(a) are spread relatively uniformly across the space. That is, the user does not appear to have a preference for specific minute-second combinations. In contrast, the user depicted in the right panel tweets at very specific times. While this program/bot- controlled PDA is not able to get the tweet out at exactly the same second each hour, the temporal distribution is clearly non-uniform.



(a) PCA (b) PDA (c) PCA (d) PDA

Fig. 3. Time distribution of tweets published by PCA and PDA

Consider the two-dimensional time space shown in Figure 3(a) and 3(b) where the x-axis is the exact minute (0-59) within the hour that the tweet was published, and the y-axis is the exact second (0-59) of that minute. Tweets' timestamp information can be used to locate each tweet as a point in this space. For each account, we count the number of tweets within each section of the grid and compute the sum of the difference between the observed frequency and the expected uniform frequency to obtain a temporal feature. Formally, let g denote the total number of grids and each time stamp can be converted to a g-dimensional vector $\boldsymbol{x} = (x_1, \cdots, x_g)$, where $x_i\{0, 1\}$ and $\sum_{i=1}^{g} x_i = 1$. This vector indicates which grid the time stamp belongs to. Suppose there are N tweets and the expected number of tweets falling in each grid should be N/g for a uniform distribution. We define a *time uniformity* metric du to measure the difference between the observed time distribution and a uniform one.

$$\boldsymbol{Y} = \sum_{i=1}^{N} \boldsymbol{x}_i - \frac{N}{g} \cdot \boldsymbol{I} \quad du = \frac{\boldsymbol{Y} \cdot \boldsymbol{Y}^T}{N/g} \tag{1}$$

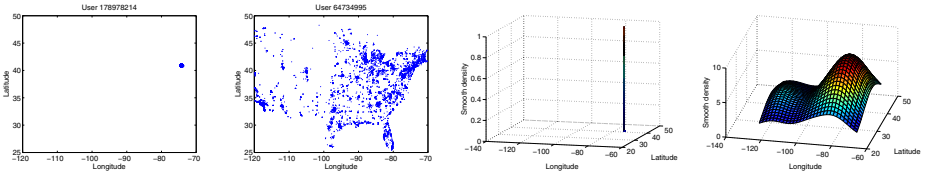
where $\boldsymbol{I} = (\underbrace{1, \cdots, 1}_{g})$ denotes a g-dimensional unit vector.

The lower value of du suggests a higher probability of uniform distribution. As can be seen in Figure 3(c) and 3(d) the distribution of du for PCAs satisfies

a log-Gaussian distribution centered around 6, while the PDAs *du* are skewed from 6 upwards.

2.2 Spatial Feature

PCA's and PDA's tweets may also exhibit different spatial distributions. As people go about their daily lives, they often tend to move around within a limited area, periodically switch between previously visited locations (e.g., home and work), and are constrained by the physical parameters bounding how fast they can travel between locations [8,20,5,4]. In contrast, PDAs, by their very nature, are not constrained. Twitter APIs can be used to tweet from multiple locations simultaneously and/or purposively designate the geo-locations that should be attached to each tweet. Figure 4(a) and 4(b) show the footprints of geo-located tweets respectively published by a PCA and a PDA. It can be seen that the PCA tweets from a small area (303.0233 km^2) in New York, while the PDA tweets from all across the United States (about 9.6302×10^6 km^2). The narrow and sharp peak in Figure 4(c) indicates a PCA visits the same locations repeatedly. In contrast, the density distribution of a PDA in Figure 4(d) is much flatter, indicating that this account rarely tweets from the same locations.



(a) PCA footprints (b) PDA footprints (c) PCA smoothed density (d) PDA smoothed density

Fig. 4. Spatial pattern of tweets published by a PCA and a PDA. The x-axis is the longitude and the y-axis is the latitude. In Figure 4(c) and Figure 4(d), the z-axis represents the smoothed frequency of visits.

We define two metrics, namely *Mobility Area (MA)* and *Unit Mobility Entropy (UME)* to capture the spatial features. MA is a measure of an account's mobility range. For a set of points in geographic space $\langle p_1, \cdots, p_n \rangle$, where $p_i = (x_i, y_i)$ consists of a longitude x_i and a latitude y_i, we can find a minimum bounding box $\langle p_{min} = (x_{min}, y_{min}), p_{max} = (x_{max}, y_{max}) \rangle$ that covers all points. MA is defined as the surface area of the bounding box in the earth.

$$\begin{aligned} MA &= \texttt{Area}(p_{min}, p_{max}) = \int_{x_{min}}^{x_{max}} \int_{y_{min}}^{y_{max}} R^2 \cdot \cos(x) dy dx \\ &= R^2(y_{max} - y_{min})(\sin(x_{max}) - \sin(x_{min})) = R^2 \triangle y \triangle \sin(x) \end{aligned} \tag{2}$$

where R is constant representing the radius of the earth.

UME measures the diversity of spatial locations visited during a specific unit of time. A smaller value indicates a higher probability of revisiting the same location. Formally, given a unique set of locations $\langle p_1, \cdots, p_n \rangle$ that appear in one's tweets and a minimum bounding box (p_{min}, p_{max}) that covers these points, the UME is defined in Equation (3).

$$UME = \frac{\sum_{i=1} \frac{f_i}{\sum_{j=1}^{n} f_j} \log \frac{\sum_{j=1}^{n} f_j}{f_i}}{\Delta T} \tag{3}$$

where f_i represents the frequency of tweets that contains the geographical point p_i and ΔT is the time interval between the earliest tweet and the most recent one.

Furthermore, we can calculate the "moving" speed of the account by checking the time stamp and geo-coding of its successive tweets. For some PDAs, where account holders may publish tweets from multiple, distant locations within a short interval, moving speed may be quite large. In contrast, PCAs are bounded by the physical constraints on human mobility.

2.3 Textual Feature

Content of PDA's tweets may also differ from that of PCA's. Given that PDAs' main objective is to disseminate a specific kind of information, they may reuse particular words. In contrast, PCAs tend to share a more diverse set of information and thus use a wider variety of words. Here we define a metric *tweet coverage* of a word as the proportion of tweets that contain the word.

We focus on two textual features: word-usage size and tweet coverage. The former measures the number of unique words appearing in an account's tweets. Since tweets of PDA aim to propagate one particular type of information, the word set is constrained towards a specific topic. In this case, the size of word set is relatively small compared to that of PCAs.

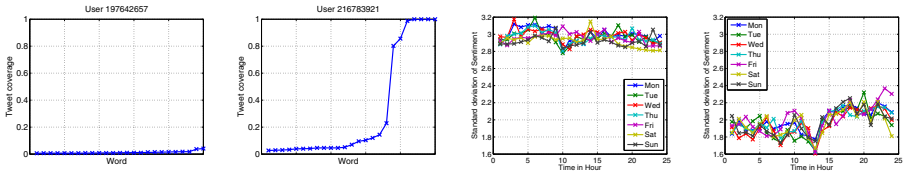
Formally, let $W = \langle w_1, \cdots, w_n \rangle$ denote the global word set and f_i^u denote the number of user u's tweets that contain the word w_i. For N posted tweets of an account, the word-usage size of the user ws^u is defined in Equation (4).

$$ws^u = \sum_{i=1}^{n} \mathbf{1}_{f_i^u \neq 0} \tag{4}$$

Tweet coverage is the probability of a single word appearing in the tweet. Given a user u's N tweets, the tweet coverage for a word w_i can be computed by $\frac{f_i^u}{N}$. Particularly, we are focused on the mean of top-k ($k \leq n$) words (referred to as top-k mean) and tweet-coverage variance of all words (referred to as global variance) for that user. Suppose we sort the words in a non-ascending order based on their tweet frequency, i.e., $\forall i, j \in [1, ws^u]$, we have $f_i^u \leq f_j^u \Leftrightarrow i \geq j$. The top-$k$ mean μ_u and global variance σ_u^2 of tweet coverage for the user account u are defined in Equation (5).

$$\mu_u = \frac{\sum_{i=1}^{k} f_i^u}{N \cdot k} \quad \sigma_u^2 = \frac{\sum_{i=1}^{n} (\frac{f_i^u}{N} - \mu)^2}{ws^u} \tag{5}$$

Figure 5(a) and 5(b) show the words' tweet coverage of 1,000 randomly sampled tweets for a PCA and a PDA. It can be easily seen that the tweet coverage for a PCA is quite low and the maximal one is about 0.04, i.e., the most frequent word appears in 4% of her published tweets. In contrast, the PDA (Figure 5(b)) uses some words in almost every tweet. In this example the PDA is a corporate account that tweets jobs for mobile phone retail in different US cities. The words with 100% tweet coverage are *job*, *mobile* and *retail*.



(a) PCA (b) PDA (c) PCA (d) PDA

Fig. 5. Tweet coverage and sentiment distribution for PCA and PDA

Moreover, since PCA's tweets are a reflection of their daily life, the sentiment of the tweets are more likely to fluctuate than that of PDAs. By adopting the lexicon for word-sentiment in existing works [16,17], we estimate a sentiment score for each tweet. Figure 5(c) and 5(d) show the standard deviation of PCA's and PDA's sentiment in Tweets covering a 24 hour time period. It can be seen that on average the PDA's tweets display less fluctuation of sentiment than the PCA's. Therefore, the deviation of sentiment is also extracted as a feature.

3 Detection Model

In this section we describe details of our detection model, including model development, parameter learning and detection function. Specifically to fit the temporal, spatial and textual features of PCAs, we propose a generative model that is adapted for stream training data. Classification of PDAs is solved by detecting the outliers of the fitted model.

3.1 Model Development

Without loss of generality, let $\mathbf{D} = \langle x_1, \cdots, x_n \rangle$ denote the values of extracted features. Here each element x_i represents the feature value of an account. The semantics depends on the feature types. For instance, x_i could indicate the log-value of du (see Equation (1) for definition) for a temporal feature, or ma (Equation (2)) for a spatial feature, or ws (Equation (4)) for a textual feature. Based on maximum-likelihood theory, to learn the model parameter μ, λ, we need to maximize the probability $Pr(\mu, \lambda|\mathbf{D})$. Using Bayesian inference, $Pr(\mu, \lambda|\mathbf{D}) = Pr(\mathbf{D}|\mu, \lambda)Pr(\mu, \lambda)$, where the $Pr(\mu, \lambda)$ is the *prior distribution*.

Under the assumption that the data $\mathbf{D}$ is generated by some Gaussian distribution $\mathcal{N}(\mu, \lambda^{-2})$, we can write the probability as in Equation (6).

$$\begin{aligned} Pr(\mathbf{D}|\mu,\lambda) &= \left(\frac{\lambda}{2\pi}\right)^{-\frac{n}{2}} exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\ &= \frac{1}{\sqrt{2\pi}}\left[\lambda^{1/2}exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^n exp\left\{\lambda\mu\sum_{i=1}^{n}x_i - \frac{\lambda}{2}\sum_{i=1}^{n}x_i^2\right\} \end{aligned} \tag{6}$$

Furthermore, since $Pr(\mu,\lambda) = Pr(\mu|\lambda)Pr(\lambda)$, we use a Gaussian and Gamma distribution as the conjugate prior distribution $Pr(\mu|\lambda)$ and $Pr(\lambda)$, as shown in Equation (7) and (8).

$$Pr(\mu|\lambda) = \mathcal{N}(\mu;\mu_0,(\alpha\lambda)^{-1}) = \sqrt{\frac{\alpha\lambda}{2\pi}}exp\left\{-\frac{\alpha\lambda}{2}(\mu-\mu_0)^2\right\} \tag{7}$$

where μ_0, α are prior distribution parameters for μ.

$$Pr(\lambda) = \mathcal{G}(\lambda;a,b) = \frac{1}{(a-1)!}b^a\lambda^{a-1}exp(-b\lambda) \tag{8}$$

Therefore, the prior distribution is represented by a product of a Gaussian and a Gamma distribution. Conversion to match the format of posterior distribution in Equation (6) gives us

$$\begin{aligned} &Pr(\mu,\lambda) = Pr(\mu|\lambda)Pr(\lambda) = \mathcal{N}(\mu;\mu_0,(\alpha\lambda)^{-1})\mathcal{G}(\lambda;a,b) \\ &\propto \left[\lambda^{1/2}exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\alpha} exp\left\{\beta\lambda\mu - \gamma\lambda\right\} \end{aligned} \tag{9}$$

Note that in Equation (9), we define for simplicity new parameters $\beta = \alpha\mu_0$, $\gamma = \frac{\alpha\mu_0^2}{2} + b$ to. Also, to maintain consistency with the posterior distribution, we constrain $a = \frac{\alpha+1}{2}$.

After unifying the posterior and prior distribution, we can represent the probability $Pr(\mu,\lambda|\mathbf{D})$ as below:

$$\begin{aligned} &Pr(\mu,\lambda|\mathbf{D}) = Pr(\mathbf{D}|\mu,\lambda)Pr(\mu|\lambda)Pr(\lambda) \propto \\ &\left[\lambda^{1/2}exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{n+\alpha} exp\left\{(\beta+\sum_{i=1}^{n}x_i)\lambda\mu - (\gamma+\frac{1}{2}\sum_{i=1}^{n}x_i^2)\lambda\right\} \end{aligned} \tag{10}$$

3.2 Model Training for Stream Data

In previous subsection we unified the prior and posterior distribution in Equation (10). Now we can define the objective function and learn the parameters by maximizing it.

Let $\theta = \langle\alpha,\beta,\gamma\rangle$ denote the parameters for prior distribution and we define the objective function as the log of the probability $Pr(\mu,\lambda|\mathbf{D})$, i.e., $\mathfrak{L}(\mu,\lambda) =$

$\log Pr(\mu, \lambda|\mathbf{D}, \theta)$. By setting partial differential $\frac{\partial \mathfrak{L}}{\partial \mu}$ and $\frac{\partial \mathfrak{L}}{\partial \lambda}$ to 0, we can estimate the value for model parameters. Without loss of generality, suppose there are two sets of training samples coming in a stream, where $\mathbf{X} = \langle x_1, \cdots, x_{n_1} \rangle$ arrives first and it is followed by $\mathbf{Y} = \langle y_1, \cdots, y_{n_2} \rangle$. Also, let (μ_x, λ_x) denote the model parameters learned purely based on $\mathbf{X}$, the sequential learning process is then illustrated in Equation (11) and (12).

$$\mu = \frac{\beta + \sum_{i=1}^{n_1} x_i + \sum_{j=1}^{n_2} y_j}{\alpha + n_1 + n_2} = \mu_x + \frac{\sum_{j=1}^{n_2} (y_j - \mu_x)}{\alpha + n_1 + n_2} \tag{11}$$

$$\lambda^{-1} = \lambda_x^{-1} + \frac{\sum_{j=1}^{n_2} [(y_j - \mu_x)^2 - \lambda_x^{-1}]}{\alpha + n_1 + n_2} - \left[\frac{\sum_{j=1}^{n_2} (y - \mu_x)}{\alpha + n_1 + n_2} \right]^2 \tag{12}$$

From the Equation (11) and Equation (12) we can see good characteristics of the model. Suppose the model has been trained based on the dataset $\mathbf{X}$ and a new dataset $\mathbf{Y}$ comes. With such sequential learning equations, instead of re-training on the whole dataset, we can simply update the model parameters with the new dataset.

3.3 Detection Function

Given the extracted features of a target account, we use the trained model to compute the probability that this account is generated by the model. The higher the value is, the more likely the account is a PCA.

Formally, suppose there is an unknown account with features $\boldsymbol{u}_0 = \langle f_1, \cdots, f_6 \rangle$. The parameters of corresponding Gaussian distribution are denoted by $\boldsymbol{M} = \langle (\cdots, (\mu_i, \lambda_i^{-1}), \cdots \rangle$. The ranking score $\boldsymbol{S}_{u_0}$ is a vector of log-likelihood that the given feature vector is generated by the model M.

$$\begin{aligned} \boldsymbol{S}_{u_0} &= \texttt{Rank}(\boldsymbol{u}_0, \boldsymbol{M}) = \langle \cdots, \log \mathcal{N}(f_i; \mu_i, \lambda_i^{-1}), \cdots \rangle \\ \log \mathcal{N}(f_i; \mu_i, \lambda_i^{-1}) &= \left(\log \lambda_i - \frac{\lambda_i}{2} (f_i - \mu_i)^2 \right) + C \end{aligned} \tag{13}$$

where C is a constant that is independent of the target account u_0 and model parameters.

The final detection is a voting process. Given the threshold vector $\boldsymbol{\delta}$, the detection function will judge whether the given account is PDA in each feature dimension. If the number of votes exceeds a threshold v_δ, the function will output 1, indicating the account is classified as PDA.

$$\texttt{Detect}(u_0, \mathbf{M}, \delta) \begin{cases} 1, & \text{if } \sum \{\texttt{Rank}(u_0, \mathbf{M}) \leq \boldsymbol{\delta} ? 1 : 0\} \geq v_\delta \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

3.4 Other Classifiers

Besides the proposed probabilistic model, we also examine the utility of other widely used classifiers in this framework. Particularly we examine Support Vector

Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree and Naive Bayes. The latter two are both implemented by Weka [10].

The SVM we adopt is developed by LIBSVM [2]. Since the number of PDA is far smaller than that of PCA, we develop three variant SVM for imbalanced classification problem. The first one over-samples minor class and is denoted as *DUP-SVM*. The second one under-samples the major class and is referred to as *RED-SVM*. Finally we increase the misclassification cost of the minority class to 100 times than that of the majority class and is referred to as *Biased-SVM*.

4 Evaluation

This section we evaluate the classifiers in terms of both effectiveness measured by F_β and efficiency measured by training time. All evaluation are based on four-fold cross-validation and average performance is reported. Then we run our generative model on the unlabeled data to mine new PDAs. The data set we use for evaluation is a collection of manually labeled accounts, 389 PDAs and 2,787 PCAs. For F_β, we set $\beta = 0.25$ because precision is relatively more important than recall in our case.

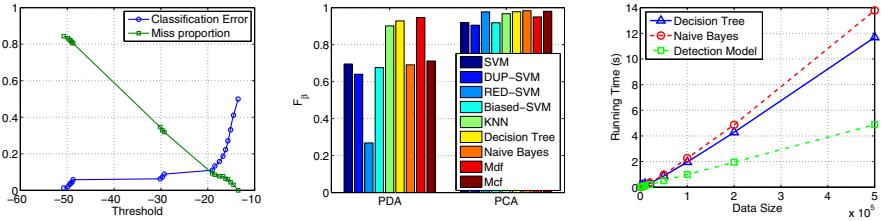
4.1 Experiments

Figure 6(a) shows the impact of threshold on the performance of the generative model. Generally, small threshold can achieve high accuracy PDA detection but may miss many PDAs. On the other hand, big threshold may reduce missing rate but lead to false identification. Figure 6(b) shows the general comparison of different methods on PDA/PCA classification. Particularly we choose the feature threshold $\boldsymbol{\delta}$ and vote threshold v_δ with regarding to maximize PDA's and PCA's F-measure, which are respectively denoted as Mdf and Mcf. The tuning process is not shown in this paper due to the page limit. We can see that our probabilistic model is either better than or close to the best performance of other classifiers. We also use synthetic data to evaluate the efficiency. Figure 6(c) shows the result and it can be seen that the time cost of training our model increases slowest as the data set grows. Note that the time cost for KNN mainly comes from classification where all training data is scanned for each classification task. The experiment shows it takes KNN 1.0334 seconds to classify one account when the training data set is 100,000. For SVM, the time cost is 401.785 seconds for 20,000 training samples in our experiment.

4.2 Exploration

In this section we run our trained model on the unlabeled data to explore new possible PDAs. By the time the paper is written, we have detected 13,871 PDAs.

Table 1 shows some of the detected PDAs. In the table, some twitter account has such symbols as **XXX** and **YYY**. They mean there is a bunch of twitter accounts with similar naming rules, where **XXX** means the type of jobs while **YYY** stands for a particular area name, e.g., tmj_tx_intern is a PDA that tweets internship in Texas, sp_arizona tweets about deals and coupons in Arizona.



(a) Threshold impact (b) Effectiveness (c) Efficiency

Fig. 6. Experiment results

Table 1. Result of exploration

Topic	Key Words	PDA	Description
job	tweetmyjobs	tmj_**XXX**_**YYY**	tweeting jobs in different areas
	intern, internship	Get**XXX**Jobs	tweeting jobs
	job, jobs	Memphiscareers	tweeting jobs in Memphis
traffic	theft, traffic	TotalTraffic**YYY**	real time traffic in **YYY** areas
	accident, police	PinellasCo911	Fire/EMS 911 Dispatches for Pinellas County, Florida
	delay	HPD_scanner	police incidents in Houston
health	healthcare, nursing	tmj_**YYY**_health	tweeting jobs of health
	hospitality, medical	tmj_**YYY**_nursing	tweeting jobs of nursing
	rescue	tofire**YYY**	fire incidents in Toronto
education	university, education	berkeleymedia	real time news in Berkeley
	instructor, news	tmj_**YYY**_edu	educational jobs in **YYY** areas
	hall	_SchoolSpring	teaching and education jobs
coupon	coupon, free	sp_**YYY**	deals and coupons in US
	service, hotel	eatcheapnearu	best restaurants with discounts
	restaurant	KidsDineFree	restaurants providing free kid-meal

5 Related Work

Two areas are related, (i) human mobility modeling, and (ii) spam detection.

As more and more people use geo-enabled smart phones to share their locations via social media, there is a large number of studies on modeling individual's mobility pattern. Generally there are two categories: (i) predicting user's location [3,11,12], and (ii) modeling continuous moving behavior [4,5,14]. These works and our work are of mutual benefit to each other. On one hand, observations of these works serve as a guideline for us to design spatial features for our detection model. On the other hand, these existing works do not differentiate common user accounts and non-human accounts. Our work can facilitate them to reduce "noise" in the data.

Many works studied the methods to battle with spammer in Twitter. In [19,18,6,22,15,21], following/followee structure was exploited.Also, spammers are usually controlled by some program and thus the times tamp of their published

tweets can be used for detection [9,23,6,13]. Since one of the spammer's motivations is to propagate some information, content-based features (e.g., ratio of URLs, number of hash tags, etc) were also used in some works [18,9,9,13].

These spammer-detection works are complementary to ours. Firstly, some features (e.g., minute-second distribution in [23]) can be used to detect PDA. Also, some spammers may disguise themselves as a PDA (e.g., adding random geo-tag in their tweets) and techniques of these works can be of great help to our framework to refine the detection result.

6 Conclusion and Future Work

The Twitter data stream is an immensely rich data resource to which many different types of entities are contributing. As such, effective use may require separation of tweets by account type. We identified types of accounts that may be especially interesting to researchers and information consumers, with specific concentration on *Public Dissemination Account (PDA)* and *Personal Communication Account (PCA)*. To separate PDAs from millions of PCAs in Twitter, we defined and extracted temporal, spatial and textual features of each account's tweets and compared our proposed probabilistic model to other conventional classifiers including SVM, KNN, Decision Tree and Naive Bayes. The experiment showed while the probabilistic model displays better or similar performance with these classifiers, it shows higher efficiency in training and is more adapted for stream data.

In future work we plan to strengthen our current system so that it can automatically build a detailed and dynamic taxonomy of PDAs, thus turning gold specks into nuggets that are more easily mined out of the Twitter stream.

References

1. Bodnar, T., Salathé, M.: Validating models for disease detection using twitter. In: WWW, pp. 699–702 (2013)
2. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Tran. on IST 2, 27:1–27:27 (2011)
3. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: CIKM, pp. 759–768 (2010)
4. Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z.: Exploring millions of footprints in location sharing services. In: ICWSM, pp. 81–88 (2011)
5. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: KDD, pp. 1082–1090 (2011)
6. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is tweeting on twitter: human, bot, or cyborg? In: ACSAC, pp. 21–30 (2010)
7. Golder, S.A., Macy, M.W.: Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science 333(6051), 1878–1881 (2011)
8. González, M.C., Hidalgo, C.A., Barabási, A.-L.: Understanding individual human mobility patterns. Nature 435, 779–782 (2008)
9. Grier, C., Thomas, K., Paxson, V., Zhang, M.: @spam: the underground on 140 characters or less. In: CCS, pp. 27–37 (2010)

10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations 11(1), 10–18 (2009)
11. Hecht, B., Hong, L., Suh, B., Chi, E.H.: Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In: CHI, pp. 237–246 (2011)
12. Kinsella, S., Murdock, V., O'Hare, N.: "i'm eating a sandwich in glasgow": modeling locations with tweets. In: SMUC, pp. 61–68 (2011)
13. Laboreiro, G., Sarmento, L., Oliveira, E.: Identifying automatic posting systems in microblogs. In: Antunes, L., Pinto, H.S. (eds.) EPIA 2011. LNCS, vol. 7026, pp. 634–648. Springer, Heidelberg (2011)
14. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: An empirical study of geographic user activity patterns in foursquare. In: ICWSM, pp. 570–573 (2011)
15. Song, J., Lee, S., Kim, J.: Spam filtering in twitter using sender-receiver relationship. In: Sommer, R., Balzarotti, D., Maier, G. (eds.) RAID 2011. LNCS, vol. 6961, pp. 301–317. Springer, Heidelberg (2011)
16. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics 37(2), 267–307 (2011)
17. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12), 2544–2558 (2011)
18. Lea, D.: Detecting spam bots in online social networking sites: A machine learning approach. In: Foresti, S., Jajodia, S. (eds.) Data and Applications Security and Privacy XXIV. LNCS, vol. 6166, pp. 335–342. Springer, Heidelberg (2010)
19. Wang, A.H.: Don't follow me - spam detection in twitter. In: SECRYPT, pp. 142–151 (2010)
20. Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabasi, A.-L.: Human mobility, social ties, and link prediction. In: KDD, pp. 1100–1108 (2011)
21. Yang, C., Harkreader, R.C., Gu, G.: Die free or live hard? Empirical evaluation and new design for fighting evolving twitter spammers. In: Sommer, R., Balzarotti, D., Maier, G. (eds.) RAID 2011. LNCS, vol. 6961, pp. 318–337. Springer, Heidelberg (2011)
22. Yardi, S., Romero, D.M., Schoenebeck, G., Boyd, D.: Detecting spam in a twitter network. First Monday 15(1) (2010)
23. Zhang, C.M., Paxson, V.: Detecting and analyzing automated activity on twitter. In: Spring, N., Riley, G.F. (eds.) PAM 2011. LNCS, vol. 6579, pp. 102–111. Springer, Heidelberg (2011)