Location Type Classification Using Tweet Content

Haibin Liu
College of IST
The Pennsylvania State University
University Park, Pennsylvania 16802
haibin@psu.edu

Bo Luo
Department of EECS
The University of Kansas
Lawrence, Kansas 66045
bluo@ku.edu

Dongwon Lee
College of IST
The Pennsylvania State University
University Park, Pennsylvania 16802
dongwon@psu.edu

***Abstract*—Location context in social media plays an important role in many applications. In addition to explicit location sharing via popular "check in" service, user-posted content could also implicitly reveals users' location context. Identifying such a location context based on content is an interesting problem because it is not only important in inferring social ties between people, but also vital for applications such as user profiling and targeted advertising. In this paper, we study the problem of location type classification using tweet content. We extend probabilistic text classification models to incorporate temporal features and user history information in terms of probabilistic priors. Experimental results show that our extensions can boost classification accuracy effectively.**

***Keywords*-classification; location detection; social media;**

I. Introduction

Social media like Twitter has become a popular platform for people to share their daily activities and statuses. People can use location-based services like Foursquare, Google Latitude, Facebook Places, etc. to "check in" at venues and share them in social media. Besides, noncheck-in tweets in Twitter can also implicitly reveal their activity-level location context even if they do not explicitly publish it. On the one hand, activity-level location can reveal users' daily activities. We are interested in finding from tweets whether the user is working in office, dining in a restaurant, or exercising in a gym, etc. On the other hand, locations reveal further information with regard to people's behavior patterns and social interaction. For example, Figure 1 shows two sample tweets both of which talk about having dinner, however, their locations are different: the first one happens at home, the second is at a restaurant or a public event *Athletes Gala*. Taking such location context into consideration, we can infer that the first dinner is a pleasant gathering with a family member, while the second one is a fun hangout with friends.

If we group users' location context into a few predefined types according to the characteristics of their activities, e.g., home is in the type *Residences*, and restaurants are in the type *Food* or *Nightlife Spots* according to Foursquare category list [1], we can obtain a clear understanding about involved user behaviors. Furthermore, if we can predict such activity level location types from users' public social media posts, we will not only arise users' privacy concerns, but also allow potential business service providers for targeted advertising.

[1]http://aboutfoursquare.com/foursquare-categories/

Figure 1. Sample tweets that mention different activity locations

In this paper, we will study the problem of classifying location types based on content of users check-in tweets. More formally, we define our research problem as:

Definition 1 (Location Type Classification) *Given a stream of tweets $d \in D$, a fixed set of location types $C = \{c_1, c_2, \ldots, c_k\}$, and a training set S of tweets labeled with types of locations where they are posted $< d, c > \in D \times C$, we wish to learn a classifier γ that maps tweets to their context location types: $\gamma : D \rightarrow C$.* □

Many researchers have already studied the problem of revealing users' locations from tweets, and shown some promising progress [1]. Different from previous work that try to reveal city-level location from tweets, we are interested in the location context in a smaller scale, the activity level. This activity-level geographic information can be essential in many applications. To illustrate:

- Service providers can utilize activity level location information to present accurate targeted advertising.
- The location type can potentially infer social ties between people. The presumption is that different social relationships have different interaction context. For example, as illustrated in Figure 1, if a twitter user tweets about activities regarding home, it is likely for her/him to interact with the family; if we find that the tweet location is a restaurant or a party, it is natural to infer that people will socialize with friends there.
- Location type revelation can also be used in user profiling. E.g., a user who tweets about Yellowstone

National Park probably enjoys traveling, while a user who talks about beer in twitter is more likely to enjoy *nightlife*, or *food*.

- Studying location type detection can also arise people's privacy awareness. People who may not be willing to share their location information in non-check-in posts should be careful because potential location can be detected from their post content.

Our goal is to filter out informative tweets and predict the location type of each tweet using content only. More specifically, we will classify each tweet into one of the nine location categories listed by Foursquare. Our contributions in this paper include the following:

- First, we present in this paper a study of location type classification through a data set of informative location sharing tweets filtered from about 1 million check-ins.
- Second, we propose a probabilistic model to incorporate temporal features to improve classification accuracy. Accuracy by this model is improved slightly from about 47% to 49% for overall dataset. However, in some specific daily time hours, the improvement is much more significant, e.g., from 37.7% to 45.3% for tweets posted at around 0 o'clock.
- Third, we propose a personalized location type classification model by incorporating users' check-in history. The experiment results demonstrate a boost in the accuracy from 47.1% to 57% for Maximum Entropy.

The rest of this paper is organized as follows: Section II presents related work; Section III introduces baseline models; Section IV describes our proposed probabilistic models; Section V describes the process of our data collection and presents an analysis of data distribution; Section VI shows our experiment results; we conclude our paper in Section VII.

II. Related Work

Several researchers have investigated the problem of geo-location detection from tweet content [1, 2, 3, 4]. Cheng et al. [1] tackle this problem in the city level. Purely based on the tweet content, the authors propose a probability language model to automatically identify words in tweets with a local geo-scope. [4] further improves user's home location prediction quality with Gaussian Mixture Models. They also employ an unsupervised measurements to rank the local words which remove the noises effectively. The authors in [2] are instead interested in the place of interest (POI) that a tweet refers to. The authors formalize the problem by ranking a set of candidate POIs using language and time models. Temporal factors need to be considered too because POIs are quite related to time. Because the POI related tweets are so sparse that the authors have to leverage search engine to enrich their language models. [3] addresses the geo-location detection problem in tweets from a different perspective. The authors are interested in matching a tweet to a specific restaurant. The question includes two parts: first, the authors need to detect which words mean a restaurant entity; second, if multiple restaurants have the same name, the authors need to detect which one is the exact match. But none of these studies are interested in the geo-location detection in activity scale like us.

Our research is directly related to the problem of location categorization. Two recent papers address this categorization problem [5, 6]. Researchers try to find out traffic patterns of venues from user generated check-in data, and take a further step to cluster the semantically related locations from these patterns. Traffic patterns can be defined as a vector of check-in frequency over a series of time units. For example, we can define daily traffic pattern that contains 24 time units, each of which represents an hour in a day; we can also define a weekly traffic pattern that contains 70 time units in which the time unit represents one tenth of a day [5]. [6] has a similar idea, in which the authors normalize the frequency into a probability density function, and call it temporal band. [5] shows that many categories indeed display quite similar daily temporal patterns, e.g., some coffee shops have similar high traffic in morning, and restaurants are frequently checked in at dinner time. [6] also demonstrates different geographic feature types have different weekly temporal bands. With such observations, the authors try to study further clustering and classification based on these similarities. But different from these papers, we do not have abundant features regarding each venue, nor are we interested in categorizing from venues' features. We instead focus on detecting location category from tweet content.

Our work is also related to short text classification. Several researchers tried to tackle this problem from different perspectives [7, 8, 9]. Sriram et. al [7]study short text classification over tweets to help users better manage information from Twitter. Phan et. al [8] try to boost the classification accuracy by gaining external knowledge from Web search results. Notice that their classification is carried out over search snippets. Sun [9] tackles the short text classification task in an information retrieval framework. The predicted category is determined by majority vote of the top search results.

III. Location Type Classification: Baseline Methods

We aim to classify the check-in location types from tweet text content. Two commonly used text classification methods are Naive Bayes [10] and Maximum Entropy [11]. We first briefly introduce these baseline methods below.

A. Naive Bayes

If we look at each check-in tweet as a document d composed of a bag of words $w_1, w_2, \ldots, w_n$, we can compute

the posterior probability that the check-in tweet belongs to category c as

$$\begin{aligned} p(c|d) &= \frac{p(c)p(d|c)}{p(d)} \\ &= \frac{p(c)p(w_1, w_2, \ldots, w_n|c)}{p(w_1, w_2, \ldots, w_n)} \\ &\propto p(c)\prod_{i=1}^{n} p(w_i|c). \end{aligned}$$

Note that $p(c)$ is the prior probability of a specific category, defined as

$$p(c) = \frac{N_c}{N}.$$

N_c is the number of check-in tweets in category c, and N is the total number of check-in tweets in training data set. The word distribution $p(w_i|c)$ can be estimated as

$$p(w_i|c) = \frac{N(w_i, c)}{\sum\limits_{w_j \in V} N(w_j, c)}$$

where $N(w_i, c)$ is number of occurrences of word w_i from category c. The check-in tweet is assigned to the best class determined by

$$\arg\max_{c \in C} p(c) \prod_{1 \le k \le n_d} p(w_k|c).$$

B. Maximum Entropy

Different from Naive Bayes, MaxEnt estimates the conditional probability directly in an exponential form instead of joint probability:

$$p(c|d) = \frac{1}{Z(d)} exp(\sum_i \lambda_i f_i(d, c)$$

where each $f_i(d, c)$ is a feature, λ_i is a constraint parameter to be estimated, and $Z(d)$ is the normalizing factor. In text classification, features are usually initiated as

$$f_{w,c'}(d, c) = \begin{cases} 0 & \text{if} c \neq c' \\ \frac{N(d,w)}{N(d)} & \text{Otherwise,} \end{cases}$$

where $N(d, w)$ is the number of times word w occurs in tweet d, and $N(D)$ is the number words in tweet d [11].

IV. Location Type Classification: Our Proposals

We propose and explore two ideas to improve the accuracy of location type classification problem.

A. Temporal Model

In this subsection, we explore the impact of temporal features in the location type classification. Our assumption is that people prefer different activities at different time. For example, the location category of *Nightlife Spot* should be more frequently checked in at night than other time. Similarly, we expect more *Food* check-ins at meal times than early morning.

To leverage temporal impacts in our classification task, we divide all check-in tweets into 24 subgroups according to the hour of their posted time, and assign a new feature $t \in \{0, 1, \ldots, 23\}$ to every check-in. Now the classification problem becomes

$$\arg\max_{c \in C} p(c|d, t)$$

which tries to maximize the conditional probability of a check-in tweet belonging to a location category given its content and posted time.

Hourly Prior Probability: One way to use this temporal feature is to apply hourly prior probability in text classifiers. Suppose the generative process of a user checking in a venue at a specific time is as follows: she first decides what kind of this check-in should be at current time, then she decides the content of that check-in tweet. Conditional independence is presumed here. That is, the content of the check-in tweet is determined only by the check-in category. More formally, in Naive Bayes, the joint probability becomes

$$p(c, d, t) = p(t)p(c|t)p(d|c).$$

For a given tweet, its posted time is always already known, the conditional probability can be estimated as

$$p(c|d, t) \propto p(c|t)p(d|c) \propto p(c|t)\prod_{i=1}^{n} p(w_i|c)$$

where $p(c|t)$ can be estimated as

$$p(c|t) = \frac{N_{ct}}{N_t}$$

while N_t is the number of check-in tweets in hour t, N_{ct} is the number of check-in tweets belonging to category c posted in hour t. $p(c|t)$ is called the hourly prior probability.

We also apply such hourly prior probability to Maximum Entropy classifier. However, since MaxEnt estimates the conditional probability $p(c|d)$ directly, as a result, the hourly prior can be applied as

$$\begin{aligned} p(c|d, t) &\propto p(c|t)p(d|c) \\ &\propto p(c|t)\frac{p(c|d)p(d)}{p(c)} \\ &\propto \frac{p(c|t)p(c|d)}{p(c)}. \end{aligned}$$

B. Boosting with User Check-in History

Different users would apparently have different activity habits, therefore we would expect different personal check-in patterns accordingly. It is quite intuitive to guess that a student checks in more frequently at *College & University* than a white-collar worker. Therefore, simply applying a same overall prior probability for all users in tweet classification

may not be fairly accurate for everyone. In this subsection, we discuss our exploration of incorporating users personal check-in history to boost classification performance .

Like hourly prior probability, we introduce a new user factor u in our model. Assuming independence between word distribution among categories and users' personal check-in habits, we can define the joint probability here as $p(c, d, u) = p(d|c)p(c|u)p(u)$, where $p(c|u)$ can be estimated from user u's personal check-in distribution. If we can retrieve adequate history check-in tweets for u, we can estimate

$$p(c|u) = \frac{N_{cu}}{N_u}$$

where N_{cu} is the number of history check-in from user u in category c, and N_u is the total number of history check-ins from u. As we are interested in maximizing conditional probability $p(c|d, u)$, and u is already known, the classification problem can formalized as

$$\arg\max_{c \in C} p(c|d, u)$$

and

$$p(c|d, u) \propto p(d|c)p(c|u).$$

Here the probability $p(c|d, u)$ includes two parts. While the first part $p(d|c)$ is a probability estimated from check-in tweet content, the second factor $p(c|u)$ is derived from user's personal check-in history. By replacing the category prior with personal check-in prior, we take both tweet content and personal habit into consideration.

Like hourly prior, to incorporate user check-in history into MaxEnt, the formula becomes

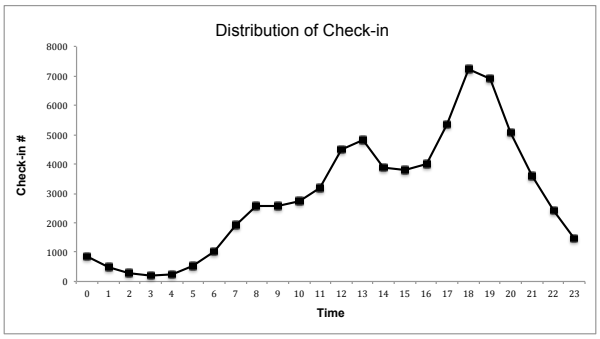
$$p(c|d, u) \propto \frac{p(c|d)p(c|u)}{p(c)}$$

where $p(c|d)$ can be estimated by MaxEnt, and $p(c)$ is the prior probability of category c in training data.

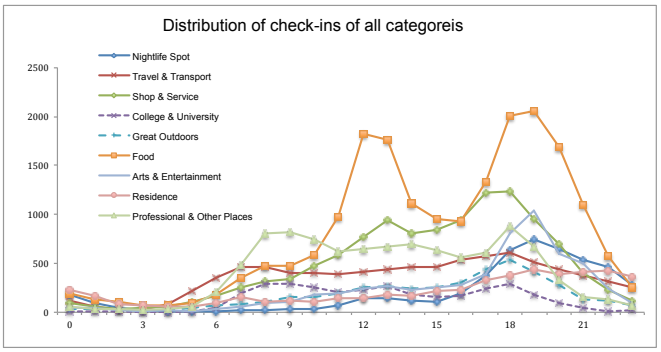
V. Experimental Setup

A. Data Collection

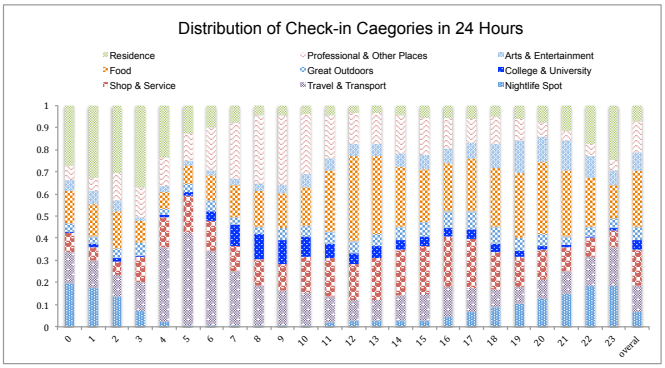
We adopt a data collection technique that relies on sampling Foursquare check-ins posted via Twiter. Using Twitter API, we search tweets with the keyword "4sq" because check-in tweets always contain URLs like http://4sq.com/xxxxxx. We monitored Twitter's public streaming API and search API for a week in May 2012, and collected about 1 million tweets, among which there are more than 220,000 foursquare check-ins. Since our focus here is to classify context location types from tweet content, we removed check-ins that only contain venue information but no user-generated comments. We also filtered tweets with less than three words. Non-English tweets were also removed from our data set. Such filtering lead to a data set of about 120,000 check-ins. The foursquare URL embedded in each check-in tweet is linked to either a "venue page" or a "check-in" page, from which we can retrieve more information about corresponding venues, and brief user profiles. Based on our best effort, we successfully tracked about 94,000 check-in tweets.



(a) # of check-ins in 24 hours



(b) Distribution of check-in for all categories



(c) Category distribution in 24 hours

Figure 2. Temporal distribution of check-ins

Foursquare has a hierarchy list of categories applied to venues, we use the top-level categories as ground truth to classify check-in tweets' location types. The top-level categories are *Arts & Entertainment, College & University, Food, Great Outdoors, Nightlife Spot, Professional & Other Places, Residence, Shop & Service, Travel & Transport*. However, there are also some venues that are not assigned to any category yet, and some venues are labeled with more than one top category. We removed such data in our

current experiments to simplify the setting. As a result, our experiment data set contains 72,643 check-in tweets with user-generated comments.

Table I
DISTRIBUTION OF CHECK-IN TWEET CATEGORIES

Category	Percentage	# of check-ins
Arts& Entertainment	8%	5781
Travel & Transport	12%	8398
Professional & Other Places	14%	10006
College & University	4%	3206
Shop & Service	16%	11661
Nightlife Spot	7%	4916
Residence	7%	5089
Food	27%	19323
Great Outdoors	6%	4263

Table I shows the distribution of check-in tweet across categories in our data set. Among the nine categories, *Food* is the most popular one (27%); *Travel & Transport (12%), Professional & Other Places (14%)*, and *Shop & Service (16%)* are less popular; the other five categories have similar percentages (around 5%) in our data.

B. Temporal Feature

To explore temporal feature's impacts, we first need to retrieve temporal information of all tweets. Because we crawled tweets from all around world, it is necessary to convert check-ins' standard UTC into local time. Such localization requires timezone information from users. Although both Twitter and Foursquare provide posted or check-in time, we find that Foursquare covers more users than Twitter, therefore we depend on Foursquare check-in API to extract tweets' localized post time.

Figure 2 shows the distribution of nine venue categories in our training data in 24 hourly time units of a day. Figure 2(a) demonstrates overall hourly check-in traffic pattern. Each point in this plot shows the number of check-in posted in an hour in our data. It shows that people check-in most frequently at 18 or 19 o'clock during a day. Figure 2(b) further illustrates detail distribution for each category. This shows us the check-in traffic changes along hours for every category. For example, we can see that *Food* are more frequently checked in at around 12 and 19 o'clock than other time of a day, and 19 is the most frequently checked in hour for *Nightlife Spot* venues. Figure 2(c) demonstrates category distribution in each hour. We also append the overall category distribution to this plot. This helps us understand not only the difference of distribution among hours, but also between each hour and the overall percentage. Compared to hours from 13-20, the category distributions in early hours like 0-8 are quite different from overall distribution. This plot also shows us which categories are the most dominant in each hour. We can see that although Table I shows that *Food* is the overall dominant category, this is not always the case in individual hours. For example, at 5 AM *Travel & Transport* venues are quite more frequently checked in than *Food*, also *Professional & Other Places* venues are more popular than any other category at 9 AM.

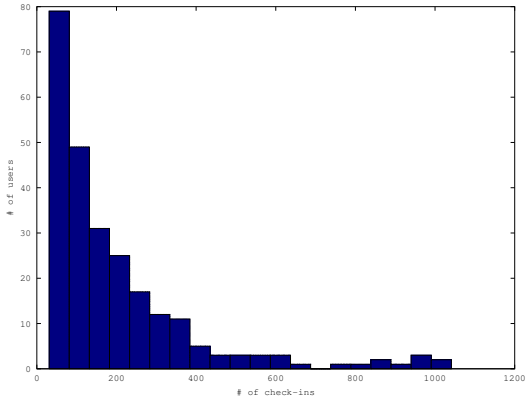


Figure 3. Distribution of user history check-in

C. User Check-in History

To evaluate the boosting impact of user check-in history in classification, we collected another data set by crawling the latest up to 1,000 tweets from randomly selected 252 users. Each of them have at least 30 check-in tweets. The total number of check-in from these users is 50,929. Distribution of the user check-in number is shown in Figure 3.

VI. EXPERIMENTAL RESULTS

The experiments are performed using Mallet toolkit [12]. Extensions are also implemented over Mallet package. The performance of all classifiers is compared in the measure of accuracy across all classes, calculated as

$$\text{accuracy} = \frac{\text{number of true positives}}{\text{number of test data set}}.$$

We use stringent five folds cross validation, and the final results are averaged over the five folds.

Figure 4 reports the results of baseline methods of Naive Bayes and MaxEnt, and also our extension of temporal model. Applying hourly prior improves overall accuracy from 47.3% to 48.6% in Naive Bayes, and from 48.6% to 49.8% for MaxEnt. We also report the details of classification performance for data in each hour in Figure 5. It shows that HourlyPrior+NB and HourlyPrior+MaxEnt achieved significant improvement in most of the hours, especially in the early hours of 0-10. This can be explained by the difference between category distribution in these hours and overall distribution as shown in Figure 2(c). Because the early hours' distribution is more different from overall distribution than other hours, the improvement is also accordingly higher by applying specific hourly prior in these hours. We also note that during the hours like 11-15, all methods have similar performance. The similar

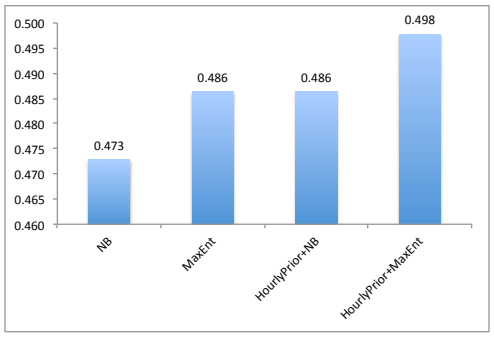


Figure 4. Overall classification accuracy

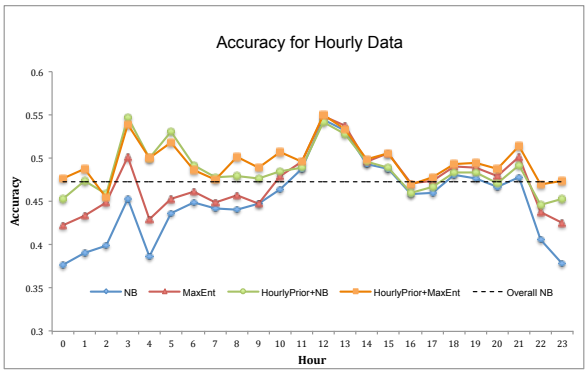


Figure 5. Classification accuracy for each hour's data

category distribution patterns during these hours with overall category distribution could also explain such classification resemblance.

Results of incorporating user history check-in are demonstrated in Figure 6. It shows that MaxEnt+UserHistory has the highest accuracy, 57.0% , compared to original 47.1% in this data set. We notice that 41.5% accuracy can be achieved using history distribution only. That is because many users are quite apt to specific venue categories. Some users may simply repeat checking in exactly the same venues. However, when they check in venues different from the history dominant category, we have to rely on tweet content for prediction.

VII. Conclusions

In this paper, we study the problem of classifying location types based on content of users' check-in tweets. We extend basic classification models by incorporating temporal features and user behavior history. The experimental results show temporal features can achieve decent performance improvement, especially in hours when the data distribution is quite different from overall daily distribution. Personal check-in history also effectively boosts the classification performance significantly.

Our future work will focus on further improving classification accuracy. One common problem in classifying

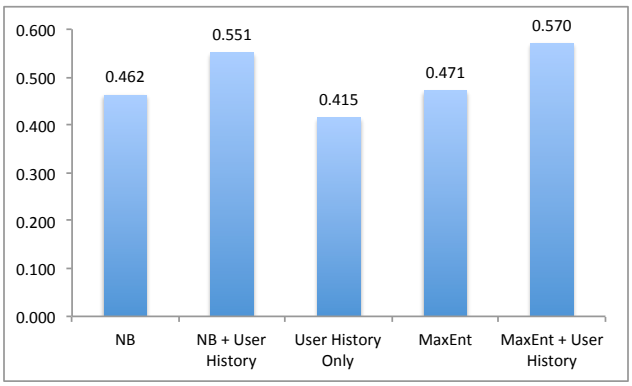


Figure 6. Classification accuracy with help of user check-in history

short text like tweets is data sparseness. To address such sparseness problem, we will study feature selection and augmentation techniques with regard to location types. Another direction is to investigate social factors. Since people interact with their friends and followers in various locations, it will be interesting to integrate social network data in this location type classification problem. We will crawl check-in data from users' friends and followers, and study the correlations between their check-in patterns. Strategies to integrate such social data into our classification framework need to be carefully studied in future.

VIII. Acknowledgments

This research was in part supported by NSF awards of DUE-0817376, DUE-0937891, and SBIR-1214331.

References

[1] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *ACM CIKM*, 2010.
[2] W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff, and M. Larson, "The where in the tweet," in *ACM CIKM*, 2011.
[3] N. Dalvi, R. Kumar, and B. Pang, "Object matching in tweets with spatial models," in *ACM WSDM*, 2012.
[4] H.-W. Chang, D. Lee, M. Eltaher, and J. Lee, "@phillies tweeting from philly? predicting twitter user locations with spatial word usage," in *IEEE/ACM ASONAM*, 2012.
[5] Z. Cheng, J. Caverlee, K. Y. Kamath, and K. Lee, "Toward traffic-driven location-based web search," in *ACM CIKM*, 2011.
[6] M. Ye, K. Janowicz, C. Mlligann, and W.-C. Lee, "What you are is when you are: the temporal dimension of feature types in location-based social networks," in *ACM SIGSPATIAL*, 2011.
[7] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *ACM SIGIR*, New York, NY, USA, 2010.
[8] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *ACM WWW*, 2008.
[9] A. Sun, "Short text classification using very few words," in *Proc. of ACM SIGIR Conference (SIGIR'12)*, 2012.
[10] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
[11] K. Nigam, "Using maximum entropy for text classification," in *In IJCAI-99 Workshop on Machine Learning for Information Filtering*, 1999, pp. 61–67.
[12] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.