DM-Group Meeting
Subhodip Biswas
10/16/2014
Papers to be discussed

1. Crowdsourcing Land Use Maps via Twitter
   Vanessa Frias-Martinez and Enrique Frias-Martinez
   in KDD 2014

2. Tracking Climate Change Opinions from Twitter Data
   Xiaoran An et al.
   Workshop on Data Science for Social Good held in conjunction with KDD 2014
Crowdsourcing Land Use Maps via Twitter

Vanessa Frias-Martinez
College of Information Studies
University of Maryland

Enrique Frias-Martinez
Telefonica Research
Madrid, Spain
Highlights

• Social media like Twitter enable individuals to generate large amounts of geolocated data that can be tapped for analysis

• The researchers think of geolocated tweets as alternative source of information for urban planning applications—characterization of landuse

• The proposed approach uses unsupervised learning to determine landuse pattern by clustering geographical regions with similar tweeting patterns
Motivations

- *Urban planners* seek to know about the utilization of the city landscape by residents
- Attempt to gather land use information through traditional approaches - questionnaire and interviews
- But here are limitations - *cost*, *willing interviewers* (mostly busy)
- *Geographic Information Systems* can be an alternative but images are not enough to capture *temporal characteristics*
- With mobile technology improvement, datasets containing information can reveal to us the interaction between *user* and *environment*
Proposal & Ideas

• Usage of *Twitter geolocated data* enabling automatic detection of landuse

• Attempt to combine *temporal* and *spatial* information of tweets (i.e. how many people may be tweeting from a particular region)

• **Besides** no access to personal information (privacy protected)

• Tries to identify all possible landuse in 2 cities- *Madrid* and *London*

• Validates predicted data with the data provided by city planning department

• Task- 1) *land segmentation* 2) *land use detection*
Land segmentation with Geolocated data

• Partitioning the land into different segment based on usage pattern

• Helps to preserve the topological properties of the tweets and preserving the geographical area under study

• This is done through Self-Organizing Maps (SOM)

• SOM has $N$ neurons organized in rectangular grid $[p, q]$ with $N = p.q$

• Any initial size $[p, q]$ can be chosen but selects the best land segmentation map that minimizes Davies-Bouldin clustering index.

• We obtain a map with each neuron referring to a region with high tweet density
Unsupervised Detection of Urban Land Uses

For each land segment $s$, a tweet-activity vector $X_s$ representing the average tweeting behavior is computed as

Step 1. An activity vector $x_{s,n}$ for land segment $s$ is built for each day $n = 1, ..., d$ in the dataset.

Step 2. Each day $n$ in the activity vector contains 72 components $x_{s,n}(t), t = 1, ..., 72$ where each one represents the number of tweets generated in land segment $s$ during a 20-minute interval $t$ in day $n$.

Step 3. An average activity vector for each land segment $s$ is computed for both weekdays $X_{s,wkd}$ and weekends $X_{s,wkn}$, each one representing the average number of tweets in land segment $s$ at each time period $t$ considering only weekdays (Monday through Friday) in the first case and weekends (Saturday and Sunday) in the second:

$X_{s,wkd}(t) = \frac{\sum_{n=1}^{d} x_{s,n}(t)}{n}$ and $X_{s,wkn}(t) = \frac{\sum_{n=1}^{d} x_{s,n}(t)}{n}$ with $n = 1, ..., d$ and $t = 1, ..., 72$.

Step 4. The final activity vector is represented as the concatenation of weekday and weekend average activity vectors $X_s = \{X_{s,wkd}, X_{s,wkn}\}$ and is normalized as $\hat{X}_s(t) = \frac{\sum_{t=1}^{72} X_{s,wkd}(t) + \sum_{t=1}^{72} X_{s,wkn}(t)}{\sum_{t=1}^{72} 1}$.

The four-step process helps to represent each land segment with a unique activity vector $X_s$ containing 144 elements representing the average weekday and weekend tweeting activity computed in 20-minute timeslots.
Unsupervised Detection of Urban Land Uses

• Use clustering over these activity vectors to automatically identify and characterize urban land areas.

• *Spectral clustering* is preferred here since
  - does not assume cluster shape
  - uses dimensionality reduction
  - easy to use based on standard linear algebra
  - low computational cost

• This technique requires
  - similarity matrix $S$ containing pairwise similarities between vectors to be clustered
  - number of clusters $k$ to compute
Evaluation of Land Uses

• The land use detection method is applied for two metropolitan areas: London and Madrid

• They are chosen since they show different density of twitter activity

• The final dataset has 49 days worth of geo-located data.

• Objective- Analyze the extent to which the land use identification algorithm detects different types of land use.
Land Segmentation and Land Uses

Figure 1: Tweeting activity signatures per cluster for London and Madrid. The Y axis represents the normalized tweeting activity and the X axis two 24-hour periods: weekdays and weekends.
Land Segmentation and Land Uses

(a) Madrid Land Uses

(b) London Land Uses

Figure 2: Physical layout of business (red), nightlife (yellow), leisure (green) and industrial (blue) clusters. Areas not marked in color are residential.
Observation

Cluster 1

- Characterized by a larger tweeting activity during weekdays than weekends.

- During weekdays the highest tweeting activity is reached at around 10:00AM and 18:30PM for London-times at which people typically get to work, go for lunch, and leave work.

- In Madrid, the signature is shifted, suggesting that working hours might happen a little bit later during the day.

- The peak of the tweeting activity during the weekends is reduced by approximately 40% when compared to weekdays.
Cluster 2

- A large difference between weekend and weekday activity (the signature is almost doubled in volume)

- During weekends, tweeting activity increases until the afternoon, and constantly decreases after that.

- Hypothesize that this cluster can be associated to Leisure or Weekend activities since users are active mostly during the weekends.

- It does not represent weekend nightlife since the tweeting activity highly decreases after 16:00PM during the weekends.
Cluster 3

- Associated to very large activity peaks at night.
- These peaks happen at around 20:00-21:00PM during weekdays and between 00:00-06:00AM during the weekends.
- The peaks happen earlier in London while a little bit later in Madrid suggesting that nightlife might continue until late hours in this city.
- Studying the physical layout of these clusters on the city maps, also suggest that this cluster might represent nightlife activities.
Observation

Cluster 4
• Signature evenly divided between weekends and weekdays

• During weekdays, there is a peak of activity in the afternoon between 6pm and 8pm.

• Activity during weekends is of the same magnitude as in weekdays.

• This is the largest cluster in terms of total area and it covers heavily residential areas in all cities.

• This type of signature represents residential land use with citizens tweeting from home at any time during the weekends and after working hours during the week.
Cluster 5
- Identified for London only.

- Its signature is characterized by a reduced activity during the weekends.

- The weekdays show a very early peak in activity (10am).

- It decreases after for the rest of the day.

- Looking at the physical layout, these clusters cover areas in the east and south of the city.

- This cluster represents Industrial land use.
Land Use validation

To validate hypothesis, evaluation results are compared against data released by

- London data store open data initiative
- Urban planning department in Madrid’s city hall

Each element \((i, j)\) in the tables represents the percentage of the official land use region that is covered by one of our land use clusters i.e., Business, Residential, Nightlife, Leisure and Industrial.
• The official Commercial and Business land uses are identified quite well by *business cluster* with area coverage between 61% – 81%.

• Similarly, the official Residential/Domestic buildings land use has a high overlap with the *residential cluster* with coverage between 56% and 68% of the official areas.

• In fact, most of the official industrial land use is subsumed by the business cluster. This might indicate that workers in the industrial areas are not using Twitter as much as people that live and/or work in that area

• The official Parks & Recreation and Greenspace & Paths land use is identified by the *leisure cluster* with overlaps between 71% and 81% of the official land use maps.
Conclusion

• An unsupervised approach for identifying land uses using location-based social media in London and Madrid.

• Results have shown that geolocated tweets can constitute a good complement for urban planners to model and understand traditional land uses.

• It can be seen as a future alternative to the traditional model of data collection from the residents as to the land usage.
Tracking Climate Change Opinions from Twitter Data

Xiaoran An
Auroop R. Ganguly
Yi Fang
Northeastern University
Steven B. Scyphers
Ann M. Hunter
Jennifer G. Dy
Highlights

• Twitter is a major repository of topical comments, and hence a potential source of information for social science research.

• Attempt to understand whether Twitter data mining can complement and supplement insights about climate change perceptions.

• A combination of techniques drawn from text mining, hierarchical sentiment analysis and time series methods is employed for this purpose.
Motivations

• Several effort have been placed on detecting public perception on climate change

• None of the previous work has utilized the widely available comment information from social network and microblogging sites.

• Conducting studies based on surveys are limited as they can only collect a limited number of participants and may also be subject to survey bias.

• Machine learning and data mining techniques to detect public sentiment on climate change, taking advantage of the freely and richly available text and opinion data from Twitter
Twitter Data

• The entire collection of data consists of 7,195,828 Twitter messages posted by users between October, 3rd 2013 and December, 12th 2013 (excluding November, 21, 22, 23 and 24).

• **Climate Change Related Twitter with Re-Tweet:** There are a total of 494,097 tweets related to climate change with 7,375 climate change tweets daily on average.

• **Climate Change Related Twitter without Re-Tweet:** The sentiment behind re-tweeted tweets is hard to detect and analyze. A total of 285,026 tweets posted in English are not re-tweeted.

• Manually labeled the Twitter data and classified them into subjective and objective groups. Within the subjective group, further distinguish them into positive and negative classes.

• Subjective tweets mean that the tweets express users' opinions or emotions regarding climate change; whereas, objective tweets are normally news regarding climate change.
Approach

• Data is treated hierarchically by first applying subjectivity detection to distinguish subjective tweets from the objective ones in the entire corpus.

• Perform sentiment analysis only within the subjective tweets.

• Text data (tweets) are pre-processed.

• We explored two classification methods for sentiment text classification.

• Naive Bayes and Support Vector Machines (SVMs) have worked well on text data.

• Feature selection is important because each tweet is typically very short (not to exceed 140 characters).

• Making a bag-of-word feature representation for each sample tweet to be very sparse.
Feature Selection

• Initially no. of features \((D)\) is 1300.

• Searching all \(2^D\) possible feature subsets is intractable.

• Rather use \textit{chi-squared metric}, which measures divergence from the distribution expected if one assumes the feature occurrence is actually independent of the class value.

\[
X^2(D, f, c) = \sum_{e_f \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{efec} - E_{efec})^2}{E_{efec}}
\]

• Higher value of \(X^2\) indicates that the hypothesis of independence is incorrect.

• Rank order the features based on this score.

• To the number of features to keep, the classification performance on a held-out validation set is measured (both macro F1 measure and accuracy as performance measures).
Sentiment Analysis

- Initially select one-fifth of entire labeled tweets randomly as validation set.
- There are 210 objective and 310 subjective tweets.
- Within the subjective tweets, there are 210 positive and 100 negative tweets.
- The rest of the four-fifth of entire labeled tweets becomes training data set which consists of 840 objective and 1190 subjective tweets, and 790 positive and 400 negative tweets.
- 10-fold cross-validation is performed on the training data set to train our model and choose the best model by comparing the performance on the validation set.
Model selection

- We perform feature selection for both SVM and Naïve-Bayes classifiers.
- We rank ordered the features based on the chi-squared scoring.
- Evaluated the performance of two classifiers for varying number of features and on both tasks, subjective vs objective and positive vs negative.
- Based on accuracy and F1-measure using 10-fold cross-validation on the training set. The results are shown to the right.
Model selection

• The performances of both algorithms vary significantly with number of features.
• As the feature size increases, serious over-fitting problem creeps in.
• It might be because of very sparse feature vectors in high dimension and limited training data size is relatively limited.
• With small number of features, the two algorithms perform well.
• A few set of candidate models are compared and tabulated.

<table>
<thead>
<tr>
<th>Table 1: Candidate Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithms</td>
</tr>
<tr>
<td>Subjectivity</td>
</tr>
<tr>
<td>Subjectivity</td>
</tr>
<tr>
<td>Subjectivity</td>
</tr>
<tr>
<td>Polarity</td>
</tr>
<tr>
<td>Polarity</td>
</tr>
<tr>
<td>Polarity</td>
</tr>
<tr>
<td>Polarity</td>
</tr>
</tbody>
</table>
Prediction and Event Detection

- With the selected subjectivity detection and sentiment polarity algorithm, the subjective tweets are extracted from the entire climate change related tweets and are divided into subgroups based on day.

- Predict the sentiment polarity on the subjective tweets as reported daily to calculate the percentage of positive and negative sentiments.

Figure 3: Subjectivity Detection and Sentiment Polarity Prediction
• The subjective and objective percentages vary largely as we move along the time axis.

• This is influenced by many factors - the news, articles published on that specific day or the occurrence of any event.

• Thus it is not easy to detect any major change or event from the subjective and objective percentages.

• It would be quite beneficial to climate sentiment studies if we can detect whether the sudden change in Twitter sentiment regarding climate change related to major climate events or extreme weather conditions.
Sentiment Polarity Percentages

- Analyze the sentiment polarity percentage trend by tracking the mean and standard deviation.

- They are calculated from a fixed-size sliding window for each time point, and plot the z-score normalization as a function of time.

- The z-score normalization can be calculated as follows:

\[ m_z = \frac{m - \bar{x}(\theta[i, \pm k])}{\sigma(\theta[i, \pm k])} \]
Sentiment Polarity Analysis

It can be assumed that an average of more than 80% of tweets believe in climate change in our data collection, this can be observed from Figure 3.

Only a small percentage of tweets that express doubt regarding climate change.

This suggests that the majority of Twitter users, studied here, think climate change is happening and believe that action is needed to mitigate it.
This paper presents result to suggest that mining social media data (Twitter accounts) can be a valuable and inexpensive way to yield insights on climate change opinions and societal response to extreme events.

It was found that major climate events can have a result in sudden change in sentiment polarity.

But considering the variation in sentiment polarity, there is still significant uncertainty in overall sentiment.

Twitter data is used to illustrate how the opinions of Twitter users can change over time and in the aftermath of specific events.

Similar approaches may be extended to other publicly available information and social media platforms.