Precise Localization of Homes and Activities: Detecting Drinking-While-Tweeting Patterns in Communities

Nabil Hossain[1], Tianran Hu[1], Roghayeh Feizi[1], Ann Marie White[2], Jiebo Luo[1] and Henry Kautz[1]

[1]Dept. Computer Science, University of Rochester, Rochester, New York, USA

[2]Dept. Psychiatry, University of Rochester, School of Medicine & Dentistry, Rochester, New York, USA

Contributions

- Fine-grained Latent Activity and Home Location Detection using Twitter data
- Applications to **Alcohol Consumption Detection**
  - fine-grained: distinguishing tweets that mention drinking alcohol vs. the user drinking alcohol vs. the user drinking alcohol at the time of tweeting
  - using 3 hierarchical SVM classifiers, with high accuracy (F-score > 0.83)
- **Home Location Prediction** (within 100 meters)
  - using SVM with accuracy > 70%, covering 71% of active users (users with at least 5 geo-tagged tweets)
  - Analyses: where drinkers live, when and where drinkers drink
- Comparison of alcohol use patterns in large city (New York City) and in suburban/rural area (Monroe County in upstate New York)

Alcohol Consumption Detection

DATASET

Millions of Geo-tagged Tweets

Keyword Filter (e.g. "drunk", "beer")

Drinking-related Geo-tagged Tweets

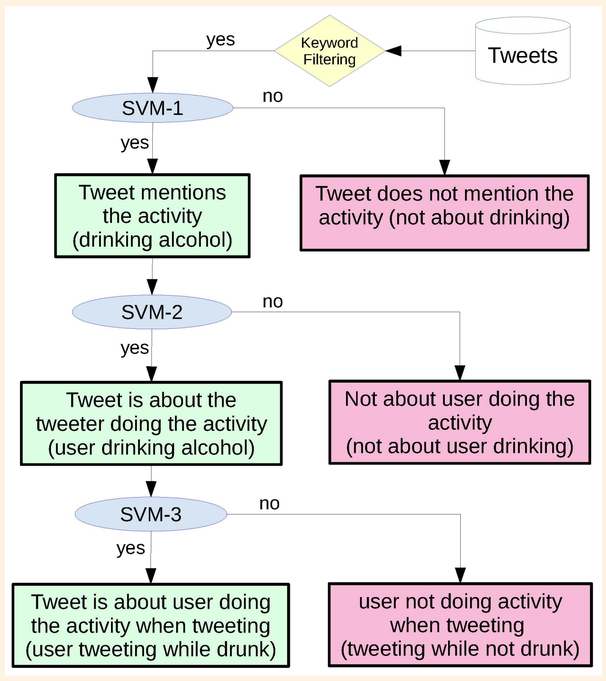
Labeling by Amazon Mechanical Turks

Alcohol Dataset

- Amazon Mechanical Turks answered 3 questions in order:
  - **Q_1: is the tweet making reference to drinking alcohol?**
  - **Q_2: if so, is the tweet about the tweeter himself drinking alcohol?**
  - **Q_3: if so, was the tweet sent when the user was drinking alcohol?**

SVM TRAINING

- Data cleanup (punctuation, url, mentions removed, text normalized)
- Trigram linguistic features (& hashtags)
  - K most frequent features in training set were used
    - K = 25 % of input data size
- Hierarchical linear SVM classifiers
  - exploit hierarchical question structure
  - 5-fold cross validation
  - F1 score for model selection
- Training data shrinks down the hierarchy
  - also restricted feature set down the hierarchy



Latent Activity Detection Flowchart

RESULTS

- Deeper questions hard to answer
- Class imbalance issues

	Q1	Q2	Q3
Class size (0, 1)	2321, 3238	579, 2044	642, 934
Precision	0.922	0.844	0.820
Recall	0.897	0.966	0.845
F-score	0.909	0.901	0.833

Alcohol Dataset Classification Results

neg. features	weights	pos. features	weights
club	-1.244	drunk	1.056
shot	-1.206	beer	1.028
party	-1.193	wine	0.998
#turnup	-0.972	alcohol	0.936
yak	-0.919	vodka	0.9
lean	-0.919	drink	0.899
crown	-0.823	tequila	0.857
root beer	-0.772	hangover	0.854
root	-0.772	drinking	0.811
wasted	-0.745	liquor	0.793
turn up	-0.673	#beer	0.779
turnup	-0.668	hammered	0.757
binge	-0.663	take shot	0.749
drunk in love	-0.593	alcoholic	0.749
in love	-0.52	get wasted	0.715
water	-0.501	champagne	0.708
turnt up	-0.499	booze	0.692
fucked up	-0.441	ciroc	0.68
fucked	-0.441	rum	0.653
water bottles	-0.423	whiskey	0.635

SVM-1 Top Features
(mentions of drinks)

neg. features	weights	pos. features	weights
she	-1.222	will	0.411
he	-0.936	when you	0.37
your	-0.87	bad	0.358
people	-0.841	when drunk	0.334
they	-0.676	with	0.318
are	-0.658	am	0.303
my mom	-0.623	get drunk	0.301
drunk people	-0.6	through	0.3
guy	-0.551	drink	0.296
#mention you	-0.5	dad	0.292
her	-0.472	us	0.286
for me	-0.454	friday	0.283
baby	-0.447	more	0.282
their	-0.431	still	0.28
his	-0.423	little	0.28
see	-0.417	drinking	0.28
most	-0.394	free	0.27
talking	-0.377	pong	0.263
the drunk	-0.368	already	0.261

SVM-2 Top Features
(pronouns, implicit drinking references)

neg. features	weights	pos. features	weights
hangover	-1.179	#url	0.662
need	-1.088	shot	0.461
want	-0.878	here	0.429
was	-0.67	#mention when	0.4
when	-0.617	bottle of wine	0.387
or	-0.605	drank	0.368
real	-0.601	now	0.36
alcoholic	-0.6	think	0.352
for	-0.561	one	0.349
last night	-0.525	good	0.327
will	-0.525	vodka	0.318
wanna	-0.523	by	0.312
tonight	-0.52	me and	0.312
got	-0.492	outside	0.307
weekend	-0.483	hammered	0.304
yesterday	-0.471	haha	0.3
was drunk	-0.47	drive	0.3

SVM-3 Top Features
(temporal references, urge to drink)

Home Location Prediction

DATASET

Millions of Geo-Tagged tweets

Keyword Filter (e.g. "home", "TV")

Remove users with less than 5 tweets

Home-related Tweets from *Active Users*

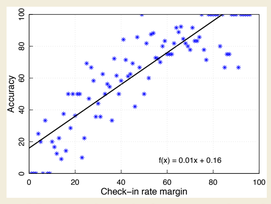
Annotate homes using Amazon Mechanical Turk

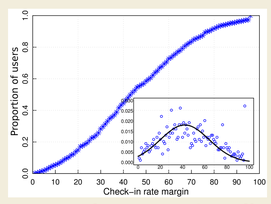
Home Dataset

Question to Turker:
- was this tweet sent from home?

SVM TRAINING

MARGIN (A, B): Difference of frequency between check-ins at location **A** and at location **B**





Most check-ins method does not always work well

SVM Features for a location:

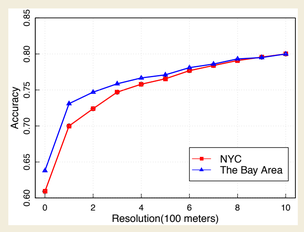
- Frequency of check-ins
- Late night check-in frequency
- Margin with next most frequent location
- Frequency as the last check-in of the day
- Distribution of check-ins over time of day
- How it behaves as origin and as destination
  - Weighted PageRank score
  - Reverse Pagerank score

RESULTS

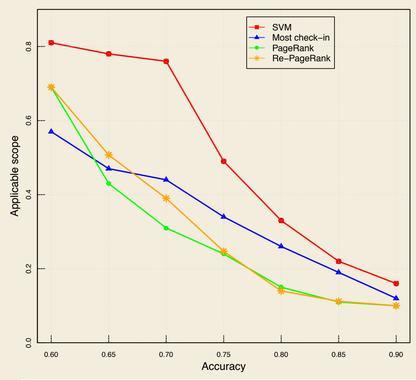
Positive Features	Weight
Check-in ratio	2.03
Margin between top two check-ins	0.19
PageRank Score	0.19
Last destination with inactive late night	0.12
Reversed PageRank score	0.09
Negative Features	**Weight**
Margin below next higher check-in	-0.30
Margin under next higher PageRank	-0.28
Margin under next higher Reversed PageRank	-0.21
Rank of Reversed PageRank	-0.07
Rank of PageRank	-0.07

SVM Top Features

Resolution vs. Accuracy Tradeoff



Resolution: Degree of granularity for home location detection



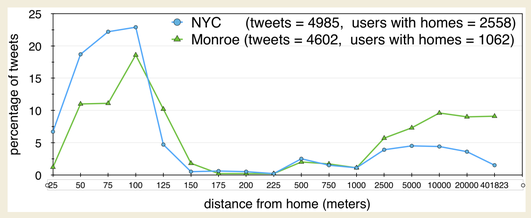
Scope (coverage): Proportion of users whose homes have been predicted

Analysis

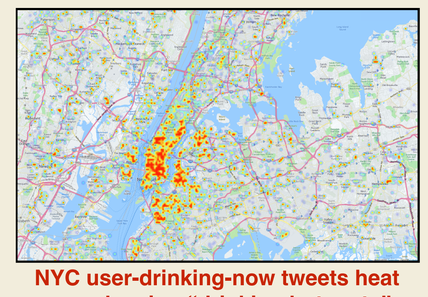
	NYC	Monroe
No. of geo-tagged tweets	1,931,662	1,537,979
Passed keyword filter	51,321	26,858
drinking-mention (Q1)	24,258	13,108
user-drinking (Q2)	23,110	12,178
user-drinking-now (Q3)	18,890	8,854
Correlation with outlet density	0.390	0.237

Classification of Drinking Related Tweets on NYC and Monroe County Datasets

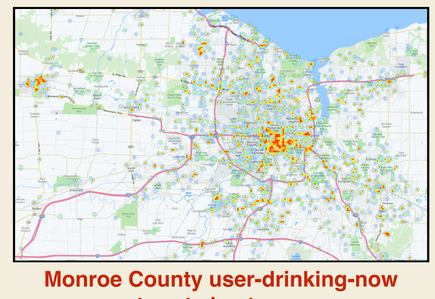
Alcohol Outlet Density: no. of businesses that serve alcohol per 100 meter grid



Histogram of Drinking Distances from Home in NYC and Monroe County



NYC user-drinking-now tweets heat map showing "drinking hotspots"



Monroe County user-drinking-now tweets heat map

Future Directions

- Explore how social interactions and peer pressure in social media influence drinking tendency
- Study user demographics and settings people go to drink-and-tweet (house, stadium, parks, etc.).
- Examine the rate of in-flow and out-flow of drinkers between neighborhoods
- Use our methods to understand other behaviors that impact community health (e.g. drug use, violence)

References

- Brownstein, J. S.; Freifeld, B. S.; and Madoff, L. C. 2009. Digital disease detection - harnessing the web for public health surveillance. N Engl J Med 260(21):2153–2157.
- Chen, M.-J.; Grube, J. W.; and Gruenewald, P. J. 2010. Community alcohol outlet density and underage drinking. Addiction 105(2):270–278.
- Hossain, N.; Hu, T.; Feizi, R.; Zheng, D.; White, A. M.; Luo, J.; and Kautz, H. 2016. Inferring fine-grained details on user activities and home location from social media: Detecting drinking-while-tweeting patterns in communities. ArchiveX.org.
- Koller, D., and Sahami, M. 1997. Hierarchically classifying documents using very few words. In Proceedings of the Fourteenth International Conference on Machine Learning (ICML).
- Lamb, A.; Paul, M. J.; and Dredze, M. 2013. Separating fact from fear: Tracking flu infections on twitter. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).
- Mahmud, J.; Nichols, J.; and Drews, C. 2012. Where is this tweet from? Inferring home locations of twitter users. In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM).
- Sadilek, A., and Kautz, H. 2013. Modeling the impact of lifestyle on health at scale. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM), 637–646.
- Scellato, S.; Noulas, A.; Lambiotte, R.; and Mascolo, C. 2011. Socio-spatial properties of online location-based social networks. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM).
- Smith, A., and Bruenner, J. 2012. Pew research centers internet & American life project: Twitter use 2012. pewinternet.org.
- Xing, W., and Ghorbani, A. 2004. Weighted Pagerank algorithm. In Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR).
- **Alcohol Dataset:** cs.rochester.edu/u/nhossain/icwsm-16-data.zip