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

Nabil Hossain¹, Tianran Hu¹, Roghayeh Feizi¹, Ann Marie White², Jiebo Luo¹ and Henry Kautz¹
¹Dept. Computer Science, University of Rochester, Rochester, New York, USA
²Dept. 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 cities (New York City) and in suburban/rural area (Monroe County in upstate New York)

Alcohol Consumption Detection

DATASET

- Millions of Geo-tagged Tweets
  - Amazon Mechanical Turks answered 3 questions in order:
    - Q1: is the tweet making reference to drinking alcohol?
    - Q2: if so, is the tweet about the tweeter himself drinking alcohol?
    - Q3: 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 the hierarchy
- also restricted feature set down the hierarchy

RESULTS

- Deeper questions hard to answer
- Class imbalance issues

SVM-1 Top Features
- (mentions of drink)
- Words: "drunk", "alcohol", "drinking"
- Root: "-0.772"

SVM-2 Top Features
- (terms related to drinking)
- Words: " whiskey", "wine", "beer"
- Root: "0.998"

SVM-3 Top Features
- (temporal reference, urge to drink)
- Words: "tonight", "last night", "now"
- Root: "-0.368"

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

Alcohol Dataset: 0.833

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
- Question to Turker: was this tweet sent from home?
- Most check-ins method does not always work well

SVM Top Features

- Resolution: Degree of granularity for home location detection
- Scope (coverage): Proportion of users whose homes have been predicted

Analysis

- Histogram of Drinking Distances from Home in NYC and Monroe County
- Alcohol Outlet Density: No. of businesses that serve alcohol per 100 meter grid
- Histogram of Drinking Related Tweets on NYC and Monroe County Datasets
- NYC user-drinking-now tweets heat map showing "drinking hotspots"

Alcohol Dataset Classification Results

- SVM Training accuracy
- SVM-1: 0.943
- SVM-2: 0.955
- SVM-3: 0.851

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

Nabil Hossain¹, Tianran Hu¹, Roghayeh Feizi¹, Ann Marie White², Jiebo Luo¹ and Henry Kautz¹
¹Dept. Computer Science, University of Rochester, Rochester, New York, USA
²Dept. 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 cities (New York City) and in suburban/rural area (Monroe County in upstate New York)

Alcohol Consumption Detection

DATASET

- Millions of Geo-tagged Tweets
  - Amazon Mechanical Turks answered 3 questions in order:
    - Q1: is the tweet making reference to drinking alcohol?
    - Q2: if so, is the tweet about the tweeter himself drinking alcohol?
    - Q3: 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 the hierarchy
- also restricted feature set down the hierarchy

RESULTS

- Deeper questions hard to answer
- Class imbalance issues

SVM-1 Top Features
- (mentions of drink)
- Words: "drunk", "alcohol", "drinking"
- Root: "-0.772"

SVM-2 Top Features
- (terms related to drinking)
- Words: " whiskey", "wine", "beer"
- Root: "0.998"

SVM-3 Top Features
- (temporal reference, urge to drink)
- Words: "tonight", "last night", "now"
- Root: "-0.368"

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