Toward Large-scale and Multi-facet Analysis of First Person Alcohol Drinking
Multi-Factorial Risk Analysis and Identification
S82

Hadi Amiri, Kara Magane, Lauren Wisk, Guergana Savova, Elissa Weitzman
Computational Health Informatics Program (CHIP)
Boston Children’s Hospital, Harvard Medical School
{firstname.lastname}@childrens.harvard.edu
NIAAA R21AA023901, NIGMS R01GM114355
Disclosure

I and my spouse/partner have no relevant relationships with commercial interests to disclose.
Learning Objectives

1. Characterize important indicators related to first-person alcohol drinking in social networks
   • Drinking context
   • Consumption level
   • Their cross correlation

2. Develop effective computational models to identify
   • First-person reports of alcohol consumption
   • Drinking context
   • Consumption level

**First person**: alcohol consumption of one or more individuals
Motivation

### Traditional Survey Data

<table>
<thead>
<tr>
<th>ID</th>
<th>Gender</th>
<th>Location</th>
<th>Generation</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>South America</td>
<td>Generation 8</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>Female</td>
<td>South America</td>
<td>Generation 5</td>
<td>1.1</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
<td>South America</td>
<td>Generation 8</td>
<td>1.44</td>
</tr>
<tr>
<td>4</td>
<td>Female</td>
<td>South America</td>
<td>Generation 5</td>
<td>1.1</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>South America</td>
<td>Generation 8</td>
<td>1.44</td>
</tr>
<tr>
<td>6</td>
<td>Female</td>
<td>Europe</td>
<td>Generation 5</td>
<td>1.44</td>
</tr>
<tr>
<td>7</td>
<td>Female</td>
<td>Europe</td>
<td>Generation 5</td>
<td>1.44</td>
</tr>
<tr>
<td>8</td>
<td>Male</td>
<td>North America</td>
<td>Generation 5</td>
<td>1.56</td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
<td>South America</td>
<td>Generation 8</td>
<td>1.56</td>
</tr>
<tr>
<td>10</td>
<td>Male</td>
<td>South America</td>
<td>Generation 8</td>
<td>1.56</td>
</tr>
<tr>
<td>11</td>
<td>Male</td>
<td>North America</td>
<td>Generation 5</td>
<td>0.999</td>
</tr>
<tr>
<td>12</td>
<td>Male</td>
<td>South America</td>
<td>Generation 8</td>
<td>1.44</td>
</tr>
<tr>
<td>13</td>
<td>Male</td>
<td>South America</td>
<td>Generation 8</td>
<td>1.44</td>
</tr>
<tr>
<td>14</td>
<td>Female</td>
<td>South America</td>
<td>Generation 5</td>
<td>0.769</td>
</tr>
<tr>
<td>15</td>
<td>Male</td>
<td>South America</td>
<td>Generation 8</td>
<td>1.56</td>
</tr>
<tr>
<td>16</td>
<td>Male</td>
<td>Europe</td>
<td>Generation 5</td>
<td>1.56</td>
</tr>
<tr>
<td>17</td>
<td>Male</td>
<td>Europe</td>
<td>Generation 5</td>
<td>1.56</td>
</tr>
<tr>
<td>18</td>
<td>Male</td>
<td>Europe</td>
<td>Generation 5</td>
<td>1.56</td>
</tr>
<tr>
<td>19</td>
<td>Male</td>
<td>North America</td>
<td>Generation 8</td>
<td>1.56</td>
</tr>
<tr>
<td>20</td>
<td>Male</td>
<td>North America</td>
<td>Generation 8</td>
<td>1.56</td>
</tr>
<tr>
<td>21</td>
<td>Male</td>
<td>North America</td>
<td>Generation 8</td>
<td>1.56</td>
</tr>
<tr>
<td>22</td>
<td>Male</td>
<td>North America</td>
<td>Generation 8</td>
<td>0.769</td>
</tr>
</tbody>
</table>

- **Challenge of traditional surveys**: low resolution, time-consuming, behavior change through time.
- **Deliver low-cost & high-resolution** views into population behaviors - first-person alcohol drinking.
Motivation

Alcohol use - a significant source of global morbidity and mortality

Alcohol use behaviors in social media
- First-person reports of alcohol use
- Temporal patterns
- Behavioral nuances

Dearth of evidence
- Drinking context
- Consumption level

Rehm et al. Addiction 2017
Liu et al. CSCW 2017
West et al. Preventive Medicine 2012
Pang et al. IEEE 2015
Outline

Dataset Development
- Crowdsourcing process
- Data analysis

Modeling
- Task and challenges
- Solution

Experiments
- Results and Analysis
Dataset Development – Crowdsourcing

1B Tweets

Full twitter firehose GPS-tagged (2014-16)

16M Tweets

Enriched Alcohol Keywords

Keyword Matched

Random Sample

10K Tweets

Crowdsourced annotation Classifier development

drinking: drink, drank, drunk, alcohol, drinks, booze, beer, etc.

First-person alcohol positive: It’s our fourth beer and we’re drinking some more!
First-person alcohol negative: I think drinking water can cure my depression!
Dataset Quality Assurance – 10K

- 10K tweets crowdsourced for annotation
  - 500 tweets annotated and injected into 10K dataset
  - Annotators maintain 80% accuracy on these tweets
  - Each tweet labeled by 3 annotators
  - Fleiss' kappa: 76.1%, substantial agreement
Dataset – 10K

4600 positive tweets
• In solitary contexts, time of drinking is often current;
• In group context, time of drinking is often past or future.

• In solitary contexts, consumption tends to be light;
• In group context, consumption tends to be heavy.

• Current drinking often indicates light consumption;
• Future and past drinking often indicate heavy consumption.
Modeling

Classification Tasks
1. First-person
2. Drinking context
3. Consumption level

Challenge
• Complexity of human language

Solution
• Linguistically-aware machine learning
Modeling

Training Data

Labels

Tweets

Feature extractor

Features

Machine Learning Algorithm

Basic Features
- Character & word ngrams
- Topic model features

Liu et al. (2017)

Enhanced Features
- Basic Features
- Syntactic and language bias features
- Sentiment & word embeddings features
- Quadratic version of these features

Classification Algorithm
- SVMs, linear kernel
- Grid search to optimize parameters
Ablation analysis showed:

1. Syntactic and
2. Word embedding

features are key source of improvement.

Figure 3: AUC classification performance.
**Analysis – 1B**

**Alcohol Cohort**: users who sent at least one alcohol-relevant tweet in 1B set.

Users in **Alcohol cohort** are considerably more active (send more tweets) than those in **Non-Alcohol cohort**.

<table>
<thead>
<tr>
<th>#Tweets</th>
<th>#Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>819.4M</td>
<td>7.3M</td>
</tr>
<tr>
<td>204.7M</td>
<td>2.4M</td>
</tr>
</tbody>
</table>

Non-Alcohol Cohort: 75.3%

Alcohol Cohort: 80.0%
Analysis – 1B

**Alcohol Cohort**: users who sent at least one alcohol-relevant tweet in 1B set.

- **#Tweets**: 819.4M for Alcohol Cohort vs. 204.7M for Non-Alcohol Cohort
- **#Users**: 2.4M for Alcohol Cohort vs. 7.3M for Non-Alcohol Cohort

Users in Alcohol cohort are considerably more active (send more tweets) than those in Non-Alcohol cohort.

**AGE, GENDER, etc.**

tweets —— users
Analysis

POS PREVALENCE

California 13.3
Texas 9.3
New York 7.0
Florida 5.4
Pennsylvania 5.1
Ohio 4.6
Illinois 4.3
Michigan 3.5
New Jersey 3.4
North Carolina 2.8
Massachusetts 2.7
Georgia 2.4
Virginia 2.4
Washington 2.1
Arizona 1.9
Maryland 1.9
Indiana 1.9
Minnesota 1.7
Wisconsin 1.6
Oregon 1.5

1st person drinking positive  1st person drinking negative
Conclusion and Future Work

Conclusion
• Effective classifiers to detect first-person report of alcohol use, context, and consumption level
• Complement traditional monitoring of alcohol use
• Public health interventions and their evaluation

Future work
• Association among alcohol use and reports of problems such as fighting, accident, vomiting, DUI

Acknowledgement: This work was supported by funding from (NIAAA R21AA023901, PI: Weitzman) and (NIGMS R01GM114355, PIs: Elhadad, Savova).
Thank you!