Natural Language Processing for Health (HLP)
April 2018 Forum
Tweet @UPennHLP #HLPMeeting

Graciela Gonzalez-Hernandez
Contact: gragon@upenn.edu
Twitter: @gracielagon

https://healthlanguageprocessing.org
Program & Speaker List

- **Welcome/Introduction** – John H. Holmes, PhD, FACE, FACMI

- **Presentations**
  - *Social Media Mining for Birth Defects Research*
    - Ari Klein, PhD – Postdoctoral Researcher, Health Language Processing (HLP) Lab, DBEI
  - *Automatic Linguistic Analysis of Descriptive Writings to Detect Abnormal Cognitive Decline and Dementia*
    - Davy Weissenbacher, PhD – Research Associate, HLP Lab, DBEI

- **Open Discussion**
Social Media Mining for Birth Defects Research

Ari Z. Klein, Abeed Sarker, Haitao Cai, Davy Weissenbacher, Graciela Gonzalez

Ari Z. Klein, Postdoctoral Researcher
Email: ariklein@pennmedicine.upenn.edu
Health Language Processing Lab (https://healthlanguageprocessing.org)
Department of Biostatistics, Epidemiology, and Informatics
Background

- Birth defects are the leading cause of infant mortality in the United States.
- The causes of the majority of birth defects remain unknown.
- Closing this knowledge gap has been challenging because sources of data for studying birth defects remain scarce.
  - Pregnant women are largely excluded from clinical trials.
  - Data from animal reproductive studies may not translate to human risk factors.
  - Pregnancy exposure registries have a variety of limitations.
Objectives

- Considering that 21% of American adults and 36% between 18-29 years of age use Twitter...
  - Assess whether Twitter could be used as a source of data for studying birth defects
  - Assess the viability of automatic NLP methods to detect sparse tweets that mention birth defects in large amounts of noisy social media data
    - clinical
    - colloquial
    - abstract
    - abbreviations
    - misspellings
    - textual representations
Methods
Results

- Inter-annotator agreement: $\kappa = 0.79$ (Cohen’s kappa)
- Recall of tweets that mention birth defects = 0.95
  - Annotated 4,169 potential false negative tweets
<table>
<thead>
<tr>
<th>Tweet</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 My son was born 5 weeks early and we found out about te fistula after.</td>
<td>+</td>
</tr>
<tr>
<td>2 My little miracle, we are so blessed to have you #hypoplasticleftheartsyndrome #hlhs</td>
<td>+</td>
</tr>
<tr>
<td>3 I'm a heart mummy - #CHD #CHDawarenessweek</td>
<td>+</td>
</tr>
<tr>
<td>4 [name] has hip clicks so we're finding out if he has hip dysplasia. I am nervous bc I had it</td>
<td>?</td>
</tr>
<tr>
<td>5 He was born with hypospadias that fixed itself so he's going to get circumsized in 2 weeks. 😞😞😞</td>
<td>?</td>
</tr>
<tr>
<td>6 Thought I'd never encounter Prune Belly syndrome. Today showed me otherwise</td>
<td>?</td>
</tr>
<tr>
<td>7 I fell asleep like 3 hours ago on this couch 😅😅. Can't watch this cleft palette infomercial though, I can't wake up depressed</td>
<td>-</td>
</tr>
<tr>
<td>8 Sally Phillips: My son has Down's syndrome - but I wouldn't want to live in a world without it via @[username]</td>
<td>-</td>
</tr>
<tr>
<td>9 Its fun with feet in #Kenya, as our #clubfoot kids run and play clubfoot free today! #runfree2030 #WorldClubfootDay</td>
<td>-</td>
</tr>
</tbody>
</table>
# Prevalence for Selected Birth Defects

<table>
<thead>
<tr>
<th>Birth Defect</th>
<th>United States</th>
<th>Social Media</th>
<th>Social Media (Adj.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anencephaly</td>
<td>1.7</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Encephalocele</td>
<td>0.8</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Holoprosencephaly</td>
<td>2.1</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Spina Bifida</td>
<td>3.5</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Congenital Cataracts</td>
<td>1.5</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Anotia/Microtia</td>
<td>1.5</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Congenital Heart Defect</td>
<td>81.4</td>
<td>11.3</td>
<td>17.5</td>
</tr>
<tr>
<td>Cleft Lip with Cleft Palate</td>
<td>5.9</td>
<td>0.8</td>
<td>1.2</td>
</tr>
<tr>
<td>Cleft Lip without Cleft Palate</td>
<td>3.2</td>
<td>1.5</td>
<td>2.4</td>
</tr>
<tr>
<td>Cleft Palate without Cleft Lip</td>
<td>6.1</td>
<td>1.3</td>
<td>2.0</td>
</tr>
<tr>
<td>Biliary Atresia</td>
<td>0.6</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Bladder Exstrophy</td>
<td>0.3</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Diaphragmatic Hernia</td>
<td>2.8</td>
<td>0.8</td>
<td>1.2</td>
</tr>
<tr>
<td>Gastrochisis</td>
<td>4.5</td>
<td>1.8</td>
<td>2.8</td>
</tr>
<tr>
<td>Limb Reduction Deformities</td>
<td>4.2</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Clubfoot</td>
<td>13.4</td>
<td>2.3</td>
<td>3.6</td>
</tr>
<tr>
<td>Craniosynostosis</td>
<td>5.0</td>
<td>1.8</td>
<td>2.8</td>
</tr>
<tr>
<td>Trisomy 13</td>
<td>1.0</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Trisomy 18</td>
<td>2.4</td>
<td>1.0</td>
<td>1.6</td>
</tr>
<tr>
<td>Trisomy 21 (Down Syndrome)</td>
<td>13.0</td>
<td>4.1</td>
<td>6.4</td>
</tr>
</tbody>
</table>

\(^a\) Per 10,000 live births  
\(^b\) Per 10,000 pregnancy outcomes, based on the estimated total number of pregnancy outcomes in the database (38,903)  
\(^c\) Per 10,000 live births, based on using the live-birth rate in the U.S. (64.58%) to adjust the 38,903 total pregnancy outcomes to 25,124—the estimated number of live births
Conclusions and Future Work

- Twitter users’ timelines offer a scarce opportunity of observing health-related information before and during pregnancies with birth defect outcomes.
- Automating the tweet classification and user timeline analyses is essential to scale the approach for epidemiological research initiatives.
Automatic Linguistic Analysis of Descriptive Writings to Detect Abnormal Cognitive Decline and Dementia

Davy Weissenbacher, Graciela Gonzalez
Detecting Abnormal Cognitive Decline

- Normal
- Cognitive Decline
- Mild Cognitive Impairment (MCI)
- Dementia

- Neuropsychological test
- Late Detection
Detecting Abnormal Cognitive Decline Earlier Using NLP

Normal → Cognitive Decline → Mild Cognitive Impairment (MCI) → Dementia

- Patient/Family Complaints
- Abnormal Linguistic Decline
- Neuropsychological test

Late Detection

Use of Natural Language Processing (NLP)
Arizona Alzheimer’s Disease Center
Longitudinal Study

Participants followed every year:

- Biological samples
- Neuropsychological tests
- **Add 1 exercise:** describing an image
Gold Standard: clinically confirmed patients free of cognitive impairment OR presenting signs of abnormal decline

Normal

A family outing at a lake shore showed people doing several things. Mom and Dad sat on a blanket while dad read a book. Dad was over comfortable without his shoes, while mom listened to the radio and poured herself a cup of coffee. Junior was having fun flying his kite. […]

Declined / Demented

Jane and Joe went out to blow. But the weather was windy in the opposite direction, so they decided to blow the joint rather than place and go home and have a bonfire in their backyard and enjoy all the cooked things they could.
Method

Collect Descriptions → Web interface → Transcriptions

Prediction

Normal In Declined

.70

.30

Machine Learning

NLP analysis

Features
- Lexical
- Semantic
- Subject

WEKA
Algorithm Selection
Feature Selection
Jane and Joe went out to blow
But the weather was windy in
the Oposit Direction, so they
decided To blow the joint
rather place and go home and
have a bond fire in Their
backyard and enjoy all the
cooked things they could ____
Jane and Joe went out to blow
But the weather was windy in the Oposit Direction, so they decided to blow the joint rather place and go home and have a bond fire in Their backyard and enjoy all the cooked things they could.

<table>
<thead>
<tr>
<th>Simple metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text length:</strong> 44 words</td>
</tr>
<tr>
<td><strong>Ratios:</strong> 10 nouns / 44 words</td>
</tr>
<tr>
<td>9 verbs / 44 words</td>
</tr>
<tr>
<td>20 tools words / 44 words</td>
</tr>
<tr>
<td><strong>Misspelling:</strong> Oposit</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stylometric metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Brunét’s index:</strong> $N^{V-0.165}$</td>
</tr>
<tr>
<td><strong>Honoré’s statistic:</strong> $100 \log N/(1-V_1/V)$</td>
</tr>
<tr>
<td><strong>Character Ngrams:</strong> ‘in th’, ‘n the’, ‘ the ‘</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>(N text length, V vocabulary)</td>
</tr>
</tbody>
</table>
Jane and Joe went out to blow but the weather was windy in the opposite direction, so they decided to blow the joint rather than go home and have a bonfire in their backyard and enjoy all the cooked things they could.

**Idea density:** heuristics estimating #claims (CPIDR 3.2)

**Word2Vec Distance:**
measure of pertinent Information, #words related to image / #words in text
## Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>False Negative</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>76.6</td>
<td>47</td>
<td>0</td>
</tr>
<tr>
<td>Bayesian Network</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- All Features</td>
<td>83.1</td>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td>- Selected Features</td>
<td>86.1</td>
<td>21</td>
<td>7</td>
</tr>
</tbody>
</table>

5 features selected: Character Ngrams, Honoré’s statistic, Misspelling ratio, Age, Word2Vec Distance

Predict ‘Normal’ for all instances

Leave-One-Out Cross Validation
Discussion

Results suggest a correlation between cognitive decline and dislocation of language ability:

- **Lexical deterioration**
  - more misspellings
  - reduced vocabulary (Honoré's statistic)

- **Semantic irregularities**
  - disgression (Character NGrams)
  - few topics described (Word2Vec distance)

[Weissenbacher et al., NAACL’16]
Future Work

Improvement and further evaluation of our linguistic instrument to detect abnormal decline
  should agree with speech language pathologists' assessment
  should be more sensitive on written descriptions than spoken descriptions
  should be more sensitive than standard linguistic exercises in neuropsychological tests