Natural Language Processing for Health (HLP)
March 2018 Gathering
Tweet @UPennHLP #HLPMeeting

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

https://healthlanguageprocessing.org
Program & Speaker List

- **Welcome/Introduction** – Graciela Gonzalez-Hernandez, PhD

- **Presentations**
  1. Aaron J. Masino, ME, PhD – Informatics Scientist, Dept of Anesthesiology and Critical Care, Dept of Biomedical & Health Informatics, CHOP
  2. Julia Parish-Morris, PhD – Research Assistant Professor of Psychology, Dept of Psychiatry, PSOM; Scientist, Center for Autism Research, CHOP
  3. Graciela Gonzalez-Hernandez, PhD – Associate Professor, HLP Lab, DBEI, PSOM

- **Open Discussion**
Identifying Abnormal Anatomy on Temporal Bone Computed Tomography Reports Using Readily Available NLP Software

Aaron J. Masino, PhD, Robert W. Grundmeier, MD, Jeffrey W. Pennington, E. Bryan Crenshaw III, PhD
Department of Biomedical & Health Informatics
The Children’s Hospital of Philadelphia
Disclosures

- Nothing to disclose
- This work was funded by the National Institutes of Deafness and Other Communication Disorders of the National Institutes of Health, project number R24 DC012207.
The Audiological & Genetic Database (AudGenDB)

- Public research database to support pediatric hearing health research

- **Contains >16,000 radiological studies**
  - Images and associated text reports
  - Nearly all studies are images of the temporal bones
  - *Studies are unlabeled*

- Desirable to enable **automated identification** of studies with **specific anatomical abnormalities**

- Minimally, **reduce** required **manual review**
Develop an Automated Labeling System
Bilateral middle and inner ear structures are unremarkable. The middle ear cavity and mastoid air cells are clear and well-aerated. The ossicular chains are unremarkable. The tegmen tympani is intact. The scutum is not eroded or blunted. The vestibular aqueduct is not enlarged bilaterally.

Accuracy < 40% for inner ear, outer ear, and mastoid regions (worse than just picking majority class); 75% for middle ear
Methods & Tools

- Python Natural Language Processing Toolkit
  - Tokenization
  - Stemming
  - Stop (common) word removal
- Python SciKit Learn
  - Feature vector generation
  - Hyper-parameter grid search
  - Model training & evaluation
- Regular expressions to replace numbers and units with fixed tokens
Data

- Corpus of 726 radiology reports
- Manually labeled with four binary labels (normal/abnormal) relative to:
  - **Inner ear**: cochlea, vestibular aqueduct, vestibular nerves, vestibules, or semicircular canals
  - **Middle ear**: tympanic membrane, ossicles, stapes, incus, malleus, or scutum
  - **Outer ear**: external auditory canal
  - **Mastoid**: mastoid regions
- Partition into stratified training (80%) and test (20%) sets
Bilateral middle and inner ear structures are unremarkable. The middle ear cavity and mastoid air cells are clear and well-aerated. The ossicular chains are unremarkable. The tegmen tympani is intact. The scutum is not eroded or blunted. The vestibular aqueduct is not enlarged bilaterally.

[75 68 17 5 80 100 48 90 12 48 16 92 0 . . . ]
Bag-of-“words” Features

- Collect word stems that appear across all training documents after removing “stop” words – length $D$.
- Feature vectors considered:
  - [1,2,3]-ngram word occurrences – binary and counts
  - [1,2,3]-ngram character occurrences – binary and counts
Bag of Words Features

- Collect the $D$ word stems n-grams that appear across all training reports.
- Convert each report to a $D$-dimensional feature vector:
  - $[1,2,3]$-ngram word and character vectors
  - Binary and ngram count vectors

<table>
<thead>
<tr>
<th></th>
<th>adjac</th>
<th>along</th>
<th>also</th>
<th>appropri</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Model Training & Evaluation

- Considered a variety of learners:
  - Naïve Bayes
  - Logistic Regression
  - SVM (Linear & Gaussian)
  - Decision Tree & Random Forest
- Grid search to span reasonable hyperparameters
- K-fold cross validation for hyperparameters and feature vector selection
- Evaluation on held-out test set
## Best Models By Region

<table>
<thead>
<tr>
<th>Region</th>
<th>Best Classifier</th>
<th>n-gram Range</th>
<th>Word / Character</th>
<th>Model Hyperparameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inner Ear</td>
<td>SVM (Linear)</td>
<td>1-3</td>
<td>Word</td>
<td>Cost parameter, $C=0.1$</td>
</tr>
<tr>
<td>Middle Ear</td>
<td>Random Forest</td>
<td>1-3</td>
<td>Character</td>
<td>Maximum Depth, $d = 5.0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Estimators (number of trees), $n_e = 200$</td>
</tr>
<tr>
<td>Outer Ear</td>
<td>Logistic Regression</td>
<td>1-2</td>
<td>Word</td>
<td>Regularization cost, $l=0.1$</td>
</tr>
<tr>
<td>Mastoid</td>
<td>SVM (Linear)</td>
<td>1-3</td>
<td>Word</td>
<td>Cost parameter, $C=0.333$</td>
</tr>
</tbody>
</table>
## Best Model Performance

<table>
<thead>
<tr>
<th>Region</th>
<th>Inner</th>
<th>Middle</th>
<th>Outer</th>
<th>Mastoid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>90 % (+16.0)</td>
<td>90 % (+30.4)</td>
<td>93 % (+7.38)</td>
<td>82 % (+19.0)</td>
</tr>
<tr>
<td>F1 Score</td>
<td>0.82</td>
<td>0.85</td>
<td>0.71</td>
<td>0.74</td>
</tr>
<tr>
<td>NPV</td>
<td>0.94</td>
<td>0.85</td>
<td>0.93</td>
<td>0.83</td>
</tr>
<tr>
<td>PPV</td>
<td>0.82</td>
<td>1.0</td>
<td>0.92</td>
<td>0.80</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.82</td>
<td>0.75</td>
<td>0.57</td>
<td>0.69</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.94</td>
<td>1.0</td>
<td>0.99</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Comparison to Baseline

<table>
<thead>
<tr>
<th>Region</th>
<th>Accuracy</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Best classifier</td>
<td>Keyword</td>
</tr>
<tr>
<td>Inner</td>
<td>90 %</td>
<td>75 %</td>
</tr>
<tr>
<td>Middle</td>
<td>90 %</td>
<td>67 %</td>
</tr>
<tr>
<td>Outer</td>
<td>93 %</td>
<td>86 %</td>
</tr>
<tr>
<td>Mastoid</td>
<td>82 %</td>
<td>65 %</td>
</tr>
</tbody>
</table>
Learning Curves

Logistic Regression

SVM Gaussian

Naive Bayes

Accuracy vs Training Examples Count
Continuing / Other Work

- Pediatric drug safety
  - Deep learning methods to identify adverse drug events in social media data
- Infant sepsis detection
  - Machine learning models incorporating EHR, vital sign, and clinical text
  - NLP: named entity recognition, sentiment analysis
Speech, language, and communication in autism

Julia Parish-Morris, PhD – Center for Autism Research, CHOP

- Autism spectrum disorder (ASD) affects ~1.5% of children (CDC, 2016)
- Center for Autism Research (CAR) established in 2008
- Use digital phenotyping (computer vision inc. 3D motion capture, wearable tech, NLP) to precisely quantify ASD from birth through adulthood
  - Improve access to screening
  - Support diagnostic decision-making
  - Develop targeted interventions
  - Measure treatment response with improved granularity
  - Revolutionize gene-brain-behavior mapping efforts

Email: parishmorrisj@chop.edu
CAR Language Group

- **Goals:**
  1. Quantify *communication across the lifespan*
  2. Identify *individual profiles* of communication strengths and weaknesses
  3. Generate machine-learning features to combine with features from computer vision and wearable technology, for screening and diagnosis

- **Data:** *Dynamic conversations* in a variety of contexts (generalizable)

- **Variable categories:**
  1. **Language** (text-based) – *social/cognitive vs. concrete focus*
  2. **Acoustic** (waveform based) – *higher and more variable Fo*
  3. **Interactive** (participant 1&2 in relation to interlocutor 1&2) – *reduced accommodation*

Email: parishmorrisj@chop.edu
Social Media Mining for Pharmacovigilance: challenges and opportunities

Case-control studies from Twitter???

Health Language Processing Lab – Penn IBI
Graciela Gonzalez-Hernandez, PhD
email: gragon@pennmedicine.upenn.edu

Funded by NLM/NIH Grant 5R01LM011176

ASCPT 2018 #Science Session – Orlando, Florida
SM data for pharmacovigilance studies

- There are about 38,220 tweets / minute about the user’s current medical conditions\(^1,2,3\)
- Patient reporting brings different perspective, more detail, info on severity and impact of ADRs in daily life. (34 studies - PMID 27558545).
- Abundant adverse event reports in SM, with a higher frequency of adverse events, particularly for ‘mild’ adverse events. (51 studies = PMID 26271492).

\(^1\)http://www.pewinternet.org/fact-sheets/health-fact-sheet/
\(^3\)http://www.internetlivestats.com/twitter-statistics/
Social Media Mining pipeline

- Data collection
- Annotation
- Classification

- Concept extraction
- Concept Mapping
- Analysis
The Aims

- **Develop and evaluate NLP methods to identify non-adherence and non-persistence and related information from Twitter data.** The methods will
  - dynamically collect a cohort of SM users that stopped taking or switched medications, did not fill a prescription, or altered their treatment,
  - extract information from the user’s *timeline* (publicly available postings over time) and *conversation threads* (postings by the user and others in reply to a posting of interest) relevant to (a) an expressed reason for these actions, (b) dosage/duration of treatment, (c) concomitant treatments, and (d) diagnosed health conditions.

- **Develop and evaluate NLP methods to identify medication use during pregnancy and pregnancy outcomes from Twitter data.**

- **Develop and evaluate methods for automatic selection of control groups to address the challenge faced when information from SM is to be used for epidemiological studies.**
Case-control study with SM data?

- Select cohort of pregnant women from SM
  - About 120 thousand, 700 million tweets
- Within that, find cases of interest
  - “Women who gave birth to a child with a birth defect and whose public tweets include tweets during pregnancy”
- Annotate (100% of the data found)
- Find matching (control) subjects
  - “Women pregnant around the same time, for whom there is no evidence that their child was born with a birth defect”

From Twitter, “I am 12 weeks pregnant”

Today, I am officially 12 weeks pregnant! Here's my first personal blog post in two years... instagram.com/p/BgoHF_--leBC/

I am 12.5 weeks pregnant and suffering terrible morning sickness all day - any recommendations on what I could take to settle it? I've tried everything :(( #help

I have a feeling I am 12 weeks pregnant because of how bloated my belly is, I can't wait to get a blood test to find out what's going on, I was supposed to have an ultrasound but didn't have it yet this month 🌟

Fast forward to this year and now here I am sitting down watching this video currently 12 weeks pregnant. Thanks for helping me smile Mark :) you have the most beautiful heart and it's comforting to know of a
From Twitter, noise

my son is 15 months and my wife is 12 weeks pregnant. when I am home it's funny dealing with his high energy and tantrums

'I am 12 Weeks Pregnant!,' Janet Mbugua Reveals She Is Expecting Baby Number Two classic105.com/i-am-12-weeks-...
Finding cases – birth defects cohort

Klein et al, 2018 (in preparation)
# Birth defects data from Social Media

Klein et al, 2018 (in preparation)

<table>
<thead>
<tr>
<th></th>
<th>Cases (n=197)</th>
<th>Controls (n=196)</th>
<th>OR or t-test [95% CI]</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median Age (IQR)</td>
<td>23 (20 to 28)</td>
<td>21 (19 to 23)</td>
<td>2 (1 to 3)</td>
<td>0.0001</td>
</tr>
<tr>
<td>Mean Age (range)</td>
<td>25 (17 to 42)</td>
<td>22 (16 to 37)</td>
<td>2.52 (1.38 to 3.66)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Women &lt;30 years</td>
<td>80% (134/168)</td>
<td>91% (129/141)</td>
<td>0.37 (0.17 to 0.77)</td>
<td>0.004</td>
</tr>
<tr>
<td>Women &lt;35 years</td>
<td>93% (156/168)</td>
<td>98% (138/141)</td>
<td>0.28 (0.05 to 1.08)</td>
<td>0.04</td>
</tr>
<tr>
<td>Missing data on age</td>
<td>14% (28/196)</td>
<td>28% (55/196)</td>
<td>0.43 (0.25 to 0.73)</td>
<td>0.0008</td>
</tr>
</tbody>
</table>

| **Race/Ethnicity**     |              |                  |                       |         |
| Caucasian              | 73% (120/164)| 55% (102/184)    | 2.19 (1.36 to 3.54)  | chi² = 23.69, d.f. = 5  P < 0.001 |
| Black                  | 13% (22/164) | 27% (51/184)     | 0.40 (0.22 to 0.72)  |         |
| Hispanic               | 9% (14/164)  | 12% (21/184)     | 0.72 (0.33 to 1.56)  |         |
| Asian                  | 2% (4/164)   | 3% (5/184)       | 0.90 (0.17 to 4.24)  |         |
| Other (Islander, Native American/Indian, Multiracial/Mixed) | 2% (4/164) | 2% (5/184) | 0.90 (0.17 to 4.24) |         |
| Missing data on race   | 16% (32/196) | 6% (12/196)      | 0.99 (1.44 to 6.58)  |         |
Social media may be particularly useful for identifying sources of intolerability that lead to non-adherence/non-persistence.

These are often not reported by physicians or patients through standard means because are considered "mild", “not serious” or are unexpected.

Significant problem, given that, on average:

- 30% of treated patients have a beneficial response
- 30% do not respond
- 10% have only side effects
- 35%-70% are non-adherent / non-persistent, often due to side-effects or perceived/real non-response
6-month persistence rate

- prostaglandin analogs 47%
- statins 56%
- bisphosphonates 56%
- oral antidiabetics 66%
- angiotensin II receptor blocker 63%
- overactive bladder medications 28%

Copyright © 2009, Academy of Managed Care Pharmacy. All rights reserved.
“I stopped taking” & “made me”

If anyone’s wondering which I doubt, the reason I stopped taking my antidepressants was because it messed with my appetite and made me feel extra drowsy and just emotionally numb. I constantly felt like a zombie, so I figured I see how I felt without them.

And well without them while in a better state of mind, I’ve gone without them before but those were my darker days aka like 2 weeks ago lol.
Thank you!

gragon@pennmedicine.upenn.edu
Twitter: @gracielagon
HLP lab (datasets and software available):
https://healthlanguageprocessing.org