JAMIA Journal Club

• May 10th, 2018
• 3:00 – 4:00 PM, EST
• Live webinar only available for CME credit
• See CME information on amia.org
How to Participate

• Speaker and Moderator converse for 40 minutes
• 20 minutes Q&A
• **Submit questions:**
  – Type into questions function box on lower right of screen
• Complete evaluation through emailed link
Disclosures

The following speakers and planners, and their life partners, have no relevant relationships with commercial interests:

Speaker: Katelynn Devinney

*JAMIA*: Michael Chiang, Jingcheng Du, Lucy Lu Wang

*AMIA*: Susanne Arnold, Pesha Rubinstein
Learning Objective

• After this live activity, the participant should be better able to:
  – Incorporate consumer social media postings as a data source for complaints foodborne illness originating from restaurants
  – Understand the performance of different text classification techniques when applied to Yelp restaurant reviews for the identification of foodborne illness complaints
Lucy Lu Wang is a PhD candidate in the Department of Biomedical Informatics and Medical Education at the University of Washington. Her research interests are biomedical knowledge representation, resource interoperability, and knowledge extraction. Her work focuses on resolving inconsistencies between structured resources, and measuring the effects of inconsistencies on secondary analysis.
Moderator: Jingcheng Du, BS

@JingchengDu

Jingcheng Du is a PhD Candidate and CPRIT Pre-doctoral Research Fellow, School of Biomedical Informatics, the University of Texas Health Science Center at Houston. His research interests are ontology, natural language processing and social media. His work focuses on developing artificial intelligence systems to analyze social media data leading to discoveries in public health issues, such as vaccination rates and mental health.
Q&A

• Type your Q into the questions function box on lower right of screen

Send your questions throughout the presentation!
JAMIA Journal Club Selection

• Effland, et al. Discovering foodborne illness in online restaurant reviews. J Am Med Inform Assoc. January 10, 2018
  – doi: 10.1093/jamia/ocx093

• Speaker:
  – Katelynn Devinney, MPH
  – New York City Department of Health and Mental Hygiene, Bureau of Communicable Disease
Author Bio

• Awardee and graduate of the Informatics-Training in Place Program (I-TIPP), a multi-agency collaboration designed for more training in public health informatics in the existing workforce while employed at a State, Tribal, Local, Territorial (STLT) health department.

• Her main research interest is to use social media for general foodborne disease surveillance and outbreak investigations.
Introduction

- Foodborne illness remains a nationwide public health concern
  - Vastly underdiagnosed and not reported to health departments
  - Outbreaks
    - Multiple illnesses associated with single source
    - Routine surveillance
    - Complaint systems
    - Majority restaurant related
    - Many undetected
- Increasing interest in the utility of social media data to identify unreported complaints and detect outbreaks of restaurant-associated foodborne illness
Introduction

• New York City (NYC)
  – 26,000 restaurants, 15,000 food retailers
  – Over 8.5 million residents; 78% report going out to eat at least once per week
  – NYC Department of Health and Mental Hygiene (DOHMH) receives ~3,500 restaurant-associated complaints submitted to the non-emergency information system, 311 and investigates ~30 outbreaks each year
• Suspected underreporting
• Outbreak investigation in 2011, identified Yelp reviews indicating foodborne illness that were not reported to 311
Introduction

• Since 2012, the Computer Science Department at Columbia University has been collaborating with DOHMH and Yelp to identify complaints of foodborne illness in Yelp reviews
  – Data mining and text classification
  – 8,523 reports
  – 10 outbreaks

• Investigate the impact of training the classifier by the prototype system
Methodology

• Yelp system design
  – Daily process, Yelp data pulled from private application program interface
  – Text classification
    • Sick
    • Multiple
    • Composite sick score
  – Reviews meeting threshold sick score qualify for DOHMH manual review; annotations are gold standard label
Methodology

• Classification Methods
  – Featurization of documents/bag-of-words
  – Prototype method – J4.8 decision tree
  – Experimental methods
    • N-grams
    • Term frequency-inverse document frequency (TF-IDF)
    • Logistic regression
    • Random forest
    • Support vector machine (SVM)
    • “Sick-Pipelined” for “Multiple
Methodology

• Enhanced data set
  – 13,526 Yelp reviews labeled by DOHMH epidemiologists since 2012
  – Balancing
    • Sick: 51% Yes, 49% No
    • Multiple: 13% Yes, 87% No
  – Split chronologically for training and evaluation (January 1, 2017) with equal class distributions
    • 11,551 training reviews
    • 1,975 evaluation reviews
Methodology

– Selection bias correction
  • All reviews in enhanced data set selected by the prototype “Sick” classifier; potentially biased
  • Added reviews labeled ‘No’ by prototype “Sick” classifier
  • Weight classification mistakes for biased and complement sampled reviews by inverses of respective probabilities for random selection from full Yelp dataset

– Training – three regimes
  • Biased
  • Gold
  • Silver
Methodology

• Evaluation
  – 1,975 biased reviews and 1,000 from complement sample
  – Four performance metrics
    • Precision
    • Recall
    • F-1 Score
    • Area under the precision-recall (AUPR) curve
  – Bias-corrected precision and recall
  – Hyperparameter tuning experiments
  – 95% confidence intervals for F-1 score and AUPR
## Important Results

- **Classifier Performance – “Sick” task**

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Regime</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score (95% CI)</th>
<th>AUPR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>J4.8</td>
<td>Prototype</td>
<td>0.48</td>
<td>0.99</td>
<td>0.65 (0.63-0.67)</td>
<td>0.83 (0.81-0.85)</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Biased</td>
<td>0.05</td>
<td>0.94</td>
<td>0.10 (0.09-0.11)</td>
<td>0.63 (0.55-0.76)</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Gold</td>
<td>0.83</td>
<td>0.88</td>
<td>0.85 (0.83-0.87)</td>
<td>0.90 (0.88-0.92)</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Silver</td>
<td>0.85</td>
<td>0.88</td>
<td>0.87 (0.85-0.88)</td>
<td>0.91 (0.90-0.93)</td>
</tr>
<tr>
<td>Random forest</td>
<td>Biased</td>
<td>0.04</td>
<td>0.91</td>
<td>0.07 (0.06-0.09)</td>
<td>0.59 (0.54-0.70)</td>
</tr>
<tr>
<td>Random forest</td>
<td>Gold</td>
<td>0.36</td>
<td>0.89</td>
<td>0.51 (0.38-0.68)</td>
<td>0.81 (0.78-0.84)</td>
</tr>
<tr>
<td>Random forest</td>
<td>Silver</td>
<td>0.70</td>
<td>0.88</td>
<td>0.78 (0.66-0.85)</td>
<td>0.87 (0.85-0.89)</td>
</tr>
<tr>
<td>SVM</td>
<td>Biased</td>
<td>0.09</td>
<td>0.95</td>
<td>0.16 (0.13-0.20)</td>
<td>0.82 (0.79-0.87)</td>
</tr>
<tr>
<td>SVM</td>
<td>Gold</td>
<td>0.33</td>
<td>0.93</td>
<td>0.49 (0.37-0.67)</td>
<td>0.88 (0.85-0.91)</td>
</tr>
<tr>
<td>SVM</td>
<td>Silver</td>
<td>0.96</td>
<td>0.74</td>
<td>0.83 (0.81-0.85)</td>
<td>0.93 (0.92-0.95)</td>
</tr>
</tbody>
</table>
Important Results

- **Classifier Performance – “Multiple” task**

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Regime</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score 95% CI</th>
<th>AUPR 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>J4.8</td>
<td>Prototype</td>
<td>&lt; 0.01</td>
<td>0.69</td>
<td>0.01</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Biased</td>
<td>0.08</td>
<td>0.56</td>
<td>0.15</td>
<td>0.25</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Gold</td>
<td>0.42</td>
<td>0.58</td>
<td>(0.09-0.26)</td>
<td>0.19-0.40</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Silver</td>
<td>0.64</td>
<td>0.58</td>
<td>(0.30-0.67)</td>
<td>0.49-0.67</td>
</tr>
<tr>
<td>Sick-Pipelined logistic regression</td>
<td>Biased</td>
<td>0.07</td>
<td>0.61</td>
<td>0.13</td>
<td>0.18</td>
</tr>
<tr>
<td>Sick-Pipelined logistic regression</td>
<td>Gold</td>
<td>0.77</td>
<td>0.56</td>
<td>(0.09-0.23)</td>
<td>0.13-0.43</td>
</tr>
<tr>
<td>Sick-Pipelined logistic regression</td>
<td>Silver</td>
<td>0.75</td>
<td>0.59</td>
<td>0.66</td>
<td>0.71</td>
</tr>
<tr>
<td>Random forest</td>
<td>Biased</td>
<td>0.04</td>
<td>0.37</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>Random forest</td>
<td>Gold</td>
<td>0.75</td>
<td>0.24</td>
<td>0.36</td>
<td>0.31</td>
</tr>
<tr>
<td>Random forest</td>
<td>Silver</td>
<td>0.74</td>
<td>0.25</td>
<td>0.37</td>
<td>0.40</td>
</tr>
<tr>
<td>SVM</td>
<td>Biased</td>
<td>0.07</td>
<td>0.65</td>
<td>0.12</td>
<td>0.18</td>
</tr>
<tr>
<td>SVM</td>
<td>Gold</td>
<td>0.35</td>
<td>0.34</td>
<td>0.35</td>
<td>0.29</td>
</tr>
<tr>
<td>SVM</td>
<td>Silver</td>
<td>0.20</td>
<td>0.30</td>
<td>0.24</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Analysis of Results

• Precision-Recall Tradeoffs
Analysis of Results

• Confusion Matrix

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Count</td>
<td>Rate (%)</td>
<td>Count</td>
<td>Rate (%)</td>
</tr>
<tr>
<td>Sick</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1882 (true negatives) 93</td>
<td>144 (false positives) 7</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>112 (false negatives) 12</td>
<td>837 (true positives) 88</td>
<td></td>
</tr>
<tr>
<td>Multiple</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>2643 (true negatives) 98</td>
<td>55 (false positives) 2</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>114 (false negatives) 42</td>
<td>163 (true positives) 58</td>
<td></td>
</tr>
</tbody>
</table>
Key Points

• Incorporation of n-grams, up to n=3 beneficial but still allowed for errors
• Logistic regression trained with the ‘Silver’ regime performed best for the ‘Sick’ task
• ‘Silver’ ‘Sick-pipelined’ logistic regression performed best for the ‘Multiple’ task
Impact

• Best performing classifier has been implemented at DOHMH and has resulted in fewer false positives, requiring fewer resources

• Future plans
  – Evaluation of this work and impact at DOHMH
  – Explore use of modern deep learning techniques
Acknowledgements

• National Science Foundation
• Alfred P. Sloan Foundation
• Fund for Public Health NYC
Q&A

• Type your Q into the questions function box on lower right of screen
Evaluation

• Complete the evaluation survey you receive in your email one hour from now
• Physicians: For live webinar only, after completing the survey, claim your CME credit through your profile at www.amia.org
Upcoming JAMIA-JC

- Today’s webinar archived at knowledge.amia.org
- The next JAMIA-JC will be on Thursday, June 14th, 2018, 3:00 – 4:00 EST
This JAMIA Journal Club has now concluded.

www.amia.org