Natural Language Processing for Health (HLP)
April 2019
Tweet @UPennHLP #HLPMeeting
Graciela Gonzalez-Hernandez
Contact: gragon@upenn.edu
Twitter: @gracielagon

https://healthlanguageprocessing.org
Program & Speaker List

Welcome/Introduction – Graciela Gonzalez-Hernandez

KEYNOTES (15 minutes + 15 minutes for questions each)

An Analysis of a Twitter Corpus for Training a Medication Intake Classifier.
Ari Klein, PhD
Post-Doctoral Researcher
Health Language Processing Lab
https://www.dbei.med.upenn.edu/post-docs/ari-klein

Social Media Science for HIV prevention: Challenges and Opportunities
Robin Stevens, PhD, MPH
Assistant Professor, School of Nursing
Director, Health Equity & Media Lab
https://www.nursing.upenn.edu/live/profiles/48-robin-c-stevens
An Analysis of a Twitter Corpus for Training a Medication Intake Classifier

Data Mining Applied to Knowledge Development

S30

Ari Z. Klein, PhD
University of Pennsylvania
March 27, 2019
Disclosure

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

After participating in this session the learner should be better able to:

- Understand the ways in which a publicly available annotated Twitter corpus can help advance the utility of social media data for observational studies in pharmacovigilance
Observational studies in pharmacovigilance have not yet harnessed social media (SM) as a potential data source.

SM data has the potential to be particularly valuable for observing medication exposure among vulnerable populations.

Leveraging SM data for observational studies requires detecting users reporting they have actually *taken* a medication mentioned in their post.

We previously developed an annotated Twitter corpus* that can be used to train machine learning algorithms to automatically detect medication intake.

* Publicly available at: https://healthlanguageprocessing.org/twitter-med-intake
Objective

• To assess how a baseline classifier trained on the general corpus—that is, on tweets that mention various types of medication—performs for specific types in the corpus
Corpus

- 27,941 tweets that mention 358 medications (including semi-automatically generated lexical variants), annotated* as:
  - Intake: indicates that the user personally took the medication, and when it was taken
    - Ex.: I've been sick for the last 3 days taking Ibuprofen just to feel better and fight swelling
    - 6363 (23%) tweets
  - Possible Intake: generally about the user’s intake, but ambiguous about whether it was actually taken and/or when it was taken
    - Ex.: I want to cry it’s that painful 😭gonna take codeine this morning for sure
    - 8214 (29%) tweets
  - No Intake: mentions a medication, but is not about the user’s intake
    - Ex.: [user name] Mine hurt for days last year!! Take some paracetamol hun 😊
    - 13,359 (48%) tweets

Classification

• Support Vector Machine (SVM) classifier
  • Radial Basis Function (RBF) kernel
  • Word n-grams (n = 1-3) as features
  • Pre-processing: lowercasing, stemming, removing URLs

• Comparing classification accuracy for overall tweets with medication types
  • 10-fold cross validation
  • Anatomical Therapeutic Chemical (ATC) categories for medication types
  • 95% confidence intervals for accuracy of tweets in each ATC group
    • Exact binomial test in R: `binom.test(x, n, p, alternative = c("two.sided"), conf.level = 0.95)`
    • x = number of successfully classified tweets in a given ATC group
    • n = number of tweets in a given ATC group
    • p = classifier’s overall accuracy (0.608)
## Classification Accuracies

<table>
<thead>
<tr>
<th>Medication Type (ATC Category)</th>
<th>Tweets</th>
<th>IAA</th>
<th>Accuracy</th>
<th>CI (95%)</th>
<th>P-value</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alimentary Tract and Metabolism</td>
<td>300</td>
<td>0.75</td>
<td>70.3%</td>
<td>64.8% - 75.4%</td>
<td>&lt;0.001</td>
<td>Greater</td>
</tr>
<tr>
<td>Anti-infectives for Systemic Use</td>
<td>219</td>
<td>0.89</td>
<td>69.9%</td>
<td>63.3% - 75.9%</td>
<td>0.01</td>
<td>Greater</td>
</tr>
<tr>
<td>Anti-neoplastic and Immunomodulating Agents</td>
<td>291</td>
<td>0.87</td>
<td>71.1%</td>
<td>65.6% - 76.3%</td>
<td>&lt;0.001</td>
<td>Greater</td>
</tr>
<tr>
<td>Blood and Blood Forming Organs</td>
<td>65</td>
<td>0.94</td>
<td>83.1%</td>
<td>71.7% - 91.2%</td>
<td>&lt;0.001</td>
<td>Greater</td>
</tr>
<tr>
<td>Cardiovascular System</td>
<td>173</td>
<td>0.92</td>
<td>80.3%</td>
<td>73.6% - 85.9%</td>
<td>&lt;0.001</td>
<td>Greater</td>
</tr>
<tr>
<td>Dermatologicals</td>
<td>1</td>
<td>1</td>
<td>0.0%</td>
<td>0.0% - 97.5%</td>
<td>0.39</td>
<td>---</td>
</tr>
<tr>
<td>Genitourinary System and Sex Hormones</td>
<td>148</td>
<td>0.89</td>
<td>73.6%</td>
<td>65.8% - 80.5%</td>
<td>0.001</td>
<td>Greater</td>
</tr>
<tr>
<td>Musculoskeletal System</td>
<td>5148</td>
<td>0.97</td>
<td>61.4%</td>
<td>60.0% - 62.7%</td>
<td>0.42</td>
<td>---</td>
</tr>
<tr>
<td>Nervous System</td>
<td>18,302</td>
<td>0.87</td>
<td>57.8%</td>
<td>57.1% - 58.5%</td>
<td>&lt;0.001</td>
<td>Less</td>
</tr>
<tr>
<td>Respiratory System</td>
<td>422</td>
<td>0.86</td>
<td>66.4%</td>
<td>61.6% - 70.8%</td>
<td>0.02</td>
<td>Greater</td>
</tr>
<tr>
<td>Sensory Organs</td>
<td>12</td>
<td>1</td>
<td>41.7%</td>
<td>15.2% - 74.8%</td>
<td>0.23</td>
<td>---</td>
</tr>
<tr>
<td>Systemic Hormonal Preparations</td>
<td>2828</td>
<td>0.95</td>
<td>73.2%</td>
<td>71.5% - 74.8%</td>
<td>&lt;0.001</td>
<td>Greater</td>
</tr>
<tr>
<td>Various</td>
<td>32</td>
<td>0.84</td>
<td>81.3%</td>
<td>63.6% - 92.8%</td>
<td>0.02</td>
<td>Greater</td>
</tr>
<tr>
<td>Overall</td>
<td>27,941</td>
<td>0.89</td>
<td>60.8%</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
General vs. Medication-Specific Training

- Tweets in relatively sparse ATC groups seem to benefit from the additional training provided by tweets that mention other types of medication.
  - A general corpus may often have utility for training machine learning algorithms to detect users reporting medication intake.

- Tweets that mention nervous system medications (e.g., Tylenol, Adderall), however, do not seem to benefit from the additional training.
  - A classifier trained specifically on tweets that mention nervous system medications may be more effective for such observational studies.
    - The classifier initially achieved an overall accuracy of 73.4% when it was trained on a subset of the corpus consisting largely of tweets that mention nervous system medications and only three of the other twelve types of medication mentioned in the larger corpus.
Implications and Future Work

• Advancing the use of SM for observational studies in pharmacovigilance
• Designing corpora for training supervised learning of health-related events
• Compare the performance of general and medication-specific training sets
Acknowledgements

• Abeed Sarker, Karen O’Connor, Graciela Gonzalez-Hernandez

• This work was funded by the NIH NLM grant number R01LM011176 and the NIH NIDA grant number R01DA046619. The content is solely the responsibility of the authors and does not necessarily represent the views of the NIH, NLM, or NIDA.
Thank you!

ariklein@pennmedicine.upenn.edu
Twitter for HIV Prevention?:
Challenges and Opportunities

Robin Stevens, PhD, MPH; PhD,
Deepthi Chittamuru, PhD,
Stephen Bonnet, BSN
Barry Slagg, &
José Bauermeister, PhD, MPH.

University of Pennsylvania
Health Equity & Media Lab

Image Credit: Ben Fearnley, “Social Media Reality”
Estimated New HIV Diagnoses in the United States for the Most-Affected Subpopulations, 2014

- Black MSM: 11,201
- White MSM: 9,008
- Hispanic/Latino MSM: 7,552
- Black Women, Heterosexual Contact: 4,654
- Black Men, Heterosexual Contact: 2,108
- Hispanic/Latina Women, Heterosexual Contact: 1,159
- White Women, Heterosexual Contact: 1,115

Categories:
- Blacks/African Americans
- Other Races/Ethnicities
HIV/AIDS in the United States

• Persistent Public Health problem disproportionately impacting MSM and ethnic/racial minorities.

• Tools on the HIV Prevention Continuum:
  • Consistent Condom Use
  • HIV Testing
  • PrEP

• Risk Behaviors
  • Condomless sex
  • Multiple partners
  • Transactional sex
  • ‘Concurrent’ sex & drug use
How can we leverage social media and online social networks to promote the pillars of the HIV prevention continuum (testing, condoms, and PrEP)?

What are the dominant topics?
You would take a bullet for your boyfriend but he wont even put on a condom when he's cheating on you...

13 likes

designersagainstaids  Birth Control  
#designersagainstaids  #DAA  #birthcontrol  
#birthcontrolpills  #condom  #sandals  #socks  
#bettersafethansomry  #keepsafecarryone
SHE BETTER BE STUDYING BABY NAMES
CAUSE I AIN'T PULLING OUT

When she asks if you pulled out and you say yea

66 likes

kingkeagz_ovo 😂😂😂😂😂
phonso_Hahahahahahahahaha
blackcoral120_Lol

34 likes

View all 9 comments
View all 5 comments
VIRUS to VIRAL STUDY:
HIV PREVENTION AMONG YOUNG MEN ON TWITTER IN THE U.S.

7,600 Relevant Tweets in 2016-2017
Methods Overview

• PrEP, Condom and Testing (PCT) related messages on Twitter from male social media users aged 13-24 years across the US

• Visual and textual HIV prevention messages were analyzed for Reasoned Action constructs and content, using manual coding and natural language processing (NLP) techniques.

• Our algorithm, which will be detailed in the presentation, positively identifies approximately 72% of relevant English U.S. tweets from 2016.
  • Include emojis & groups of words (e.g. ‘nuts’ with/without ‘lick’)

• Building out the Porn Qualifier – “We know it when we see it”

• Our sample include approx. 7,600 prevention, sexual risk related and general sex-related tweets.

• 80% inter-coder reliability on key coding constructs (4 coders).

• Manually coded all prevention/risk related tweets, 17.5% of Sex tweets in general
DIGITAL EPIDEMIOLOGY:

CONTENT ANALYSIS
all black lets go do
a hit hoodie on they
like like who is that
To Address Context Collapse:
Hybrid User/Tweet level Analysis
Preliminary Results

• 2016 Twitter - yields → 5,600 Prevention Related, No Risk, 25K Sex in general
• Manually coded 8,611 (All prevention related tweets, 17.5% of Sex tweets in general)
• Manual Coders identify 10.8% of final sample as “not sex related”
• Less than 0.3% were porn with deployment of porn classifier
• Aggressive marketing of porn via health hashtags
• Most “#high risk” words are porn seeds
• Condoms are the best word for Condoms
• 12% of tweets were replies
<table>
<thead>
<tr>
<th>TOPIC</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condoms</td>
<td>0.027</td>
<td>0.16</td>
</tr>
<tr>
<td>HIV test</td>
<td>0.134</td>
<td>0.34</td>
</tr>
<tr>
<td>PrEP</td>
<td>0.031</td>
<td>0.17</td>
</tr>
<tr>
<td>HIV</td>
<td>0.607</td>
<td>0.49</td>
</tr>
<tr>
<td>STI</td>
<td>0.087</td>
<td>0.28</td>
</tr>
<tr>
<td>Anti-risk</td>
<td>0.166</td>
<td>0.37</td>
</tr>
<tr>
<td>Pro-risk</td>
<td>0.009</td>
<td>0.10</td>
</tr>
<tr>
<td>Modeling</td>
<td>0.003</td>
<td>0.05</td>
</tr>
<tr>
<td>Norm</td>
<td>0.026</td>
<td>0.16</td>
</tr>
<tr>
<td>Multipartner</td>
<td>0.017</td>
<td>0.13</td>
</tr>
<tr>
<td>Drugs</td>
<td>0.027</td>
<td>0.16</td>
</tr>
<tr>
<td>Transactional</td>
<td>0.007</td>
<td>0.08</td>
</tr>
<tr>
<td>Unprotected</td>
<td>0.012</td>
<td>0.11</td>
</tr>
<tr>
<td>Pharma</td>
<td>0.001</td>
<td>0.03</td>
</tr>
<tr>
<td>Porn</td>
<td>0.003</td>
<td>0.06</td>
</tr>
<tr>
<td>News</td>
<td>0.503</td>
<td>0.50</td>
</tr>
</tbody>
</table>
Spread & Endorsements

Weed is illegal but knowingly giving ppl HIV isn't. White ppl make the DUMBEST decisions🤦‍♂️🤔

Favorited 2962, 2800 Retweets
Currently non-existent on Twitter
Messages Flagged by coders

• Last night Was Mad Real Terrific All I Remember Is Magnums And Bitches
• Condom sex is foreplay tbh
• You think cuz she bad she doesn't have a STD ?
• Scientists discover potential antibodies through early stage study of HIV infection..
• if ur dick wont get hard u can catch me on another dick u hear me
• DESPICABLE: Clinton Foundation Caught Watering Down AIDS Drugs to Africa
Robotic US Election Influence Messages

• RT @FeministaJones: Well then that's when the HIV kicks in; women 50+ are among the fastest growing contraction group https://t.co/jsCq8Lvp

• I want to see Hillary's fans reaction when they see her naked statues! I bet she has HIV #ReleaseClintonsMedicalRecords

• RT @TeenageCloseted: In case you didn't know, Mike Pence wanted to divert all national HIV funding into "conversion therapy" for the LGBT

• RT @WorldTruthTV: Teen Girl Infects 324 Men With HIV On Purpose | World https://t.co/telXZBWPRi https://t.co/SMzvAy3UQ3

• RT @TrumpinIndiana: EXCLUSIVE:Clinton Foundation AIDS Program Distributed Watered-Down Drugs To Third World Countries https://t.co/OlvMX9
Next Steps

• Gender filtering seems flawed
• Conduct analysis of characteristics of tweets that are most often retweeted.
• Examine the role of tweeter characteristics (followers, age, geography)
• Use annotated dataset to refine the classifier for application on older twitter data.
• Test association with HIV prevalence and other risk behaviors.
HEALTH EQUITY MEDIA LAB TEAM

• JOSÉ BAUERMEISTER, MPH, PHD
• DEEPTI CHITTAMURU, PHD
• BARRY SLAGG
• STEPHEN BONNET, BSN
• IFY OBI, MPH
• MARAWIT ABERA
• ELIZABETH LAZARUS
• SARAH HUANG
• GENEVIEVE AFRIFA
• PHIL EZEALA
• KAHAAI KENYATTA
• CIARRA LAWSON, MPH, MSN
• SALLY KONG
• YOUTH RESEARCH TEAM
• LYLE UNGAR, PHD