Big Data for Reproductive Health: Using Natural Language Processing Techniques to examine Stigma with Cervical Cancer in Kenya

Foxx Hart, Lynne Wang, Alexandra Lawrence, Neha Shrishail, Amy Finnegan, PhD[1,2,3], Megan Huchko, MD, MPH[1,2,5], Kelly Hunter[1,2,4]

1. Duke Global Health, 2. Duke Center for Global Reproductive Health, 3. IntraHealth International, 4. Stanford School of Public Policy, 5. Duke OB/GYN

Duke UNIVERSITY | BASS CONNECTIONS Global Health

IntraHealth INTERNATIONAL Because Health Workers Save Lives.

Introduction

- Natural language processing (NLP) is gaining popularity due to its cost-effectiveness and ability to handle larger datasets relative to traditional hand coding
- Qualitative hand coding is generally time intensive and expensive
- Currently there is little research on the application of NLP to stigma data. We sought to determine if NLP could be applied to stigma data

NLP Approach Used for our Analysis

- Using Latent Dirichlet Allocation (LDA), a topic modeling NLP technique, which uses the following equation and document-probability matrices to determine probability of finding a word in a topic

$$p_d(w) = \sum_k (\theta_{d,k} + \alpha)(\beta_{w,k} + \eta)$$

- Gamma optimization to determine the probability of a topic given a token

Research Question

Does LDA, when optimized according to Cv topic coherence criterion, generate the same topics (in both quantity and kind) as hand-coding for stigma interview data? If not, do they differ in quantity, kind, or both?

Hypotheses

(1) NLP will produce fewer topics whose qualitative similarity to topics from hand-coding will vary
(2) NLP will be most likely to miss or misclassify more nuanced topics, such as stigma

Data

- 26 in-depth interviews conducted among **Kenyan women (both HIV-positive and negative)**, community health volunteers (CHVs), and healthcare providers in Kisumu, Kenya in 2019

Methods

- We applied NLP to three distinct, predetermined document sizes:
  - an individual's entire interview
  - an individual's response to the entire interview question below
  - an individual's response to one segment of the interview question below

I would like to understand what you know about human papillomavirus, or HPV. What have you heard about HPV?

a. What is the difference between HPV and cervical cancer?
b. What factors put someone at risk for HPV? How do these risk factors differ from those for cervical cancer?
c. Imagine if someone you know was diagnosed with HPV. What are some of the thoughts that come to mind?
   i. What if this person were a loved one?
   i. What if you were diagnosed with HPV?
d. *If stated 'in a relationship' in demographic survey:* What does your husband/partner think about HPV? How does this compare to other men in the community? What do they know about HPV? (**1st document size)**
e. *If stated 'unmarried/single' in demographic survey:* What do men in the community think about HPV? What do they know about HPV?

Analysis Process

Data Cleaning

- Extracted chosen document size from each interview by hand
- Added interview ID and text to Microsoft excel spreadsheet
- Removed common stop words in R

Generating Topics

- Separated each response into bigrams (two-word tokens) and created Term-Document Matrix (TDM) for the document size
- Ran LDA function from "topicmodels" package on TDM
- Selected top ten terms from each topic, filtering out words common across topics

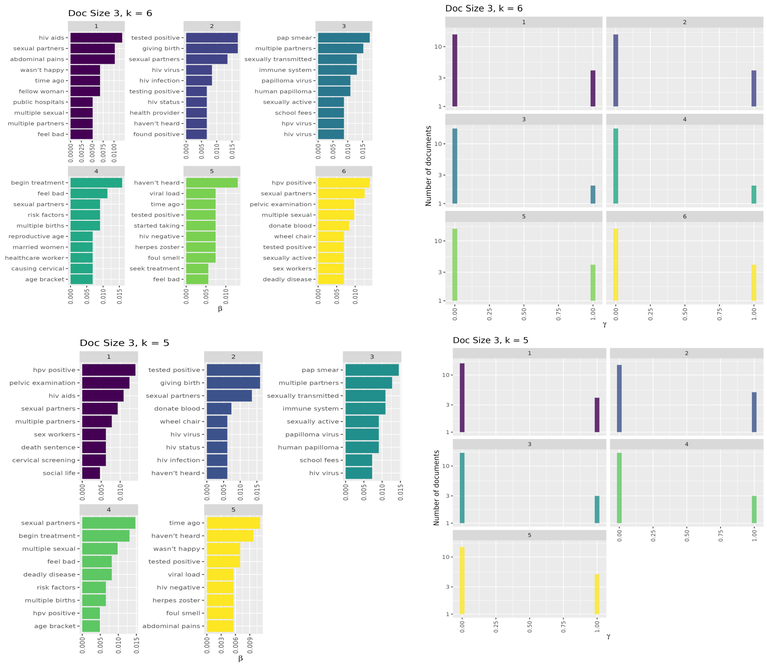
Analysis

- Plotted probabilities of each bigram belonging to a topic
- Repeated process with increasing values of k until reaching a k where at least one topic did not contain any documents
- Attempted comparison between most probable bigrams in a topic and stigma categories by hand

Limitations

- Small sample size and n-gram size limited how much nuance topic modeling could capture
- Limited foundational understanding of LDA as an algorithmic and mathematical process

Results



Topics and gamma probabilities for k=5 and k=6 for an entire interview

- Bigrams for three topics over entire interviews yielded the most easily interpretable results
- In most instances, the probability of a topic given a token was either zero or one
- Topic lists alluded to interview content, but not stigma in particular

Discussion

- Rudimentary NLP is not adequate for identifying stigma in qualitative data
- Stigma categories (i.e. enacted, anticipated, internalized) are determined by specific nuances that cannot be found in groups of bigrams

Next Steps

- Compile findings in a manuscript
- Look into applying methods on larger stigma-rich datasets, such as social media discussions
- Use additional NLP methods outside of topic modeling to draw insights on stigma in qualitative data

Guetterman et al. "Augmenting Qualitative Text Analysis with Natural Language Processing: A Methodological Study." J Med Internet Res 2018; 20(6): e231. https://www.jmir.org/2018/6/e231.
Rosner et al. "Evaluating topic coherence measures." https://arxiv.org/abs/1403.6397.
Verspoor et al. "Deploying natural Language Processing for Social Science Analysis." PDF File. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.1035&rep=rep1&type=pdf.

Acknowledgments

We would like to extend our appreciation to our team leaders (Dr. Amy Finnegan, Dr. Megan Huchko, and Kelly Hunter for their guidance and support this year. We thank the Bass Connections program, the Duke Global Health Institute, and the Duke Center for Global Reproductive Health for their funding and support.