The use text mining for clinical research: An overview of the methodology

Ross Jacobucci
University of Notre Dame

Slides

Slides posted at rjacobucci.com/presentations

Text-Based Responses

The Ubiquity of Likert Items

Critiquing the use of Likert type items isn't new: e.g., Hartley, 2014; Kyllonen & Bertlin, 2013; Subedi, 2016; Ogden & Lo, 2011; Runge, Waller, MacKenzie, & McGuire, 2014; Lazarsfeld, 1944.

In assessing suicide risk factors, Likert type dominate, which can bludgeon nuance in a participants response, as well as limit the amount of detail derived from the assessment.

Beyond Closed-Response Formats

Text has been used extensively in social media – However, can't control topics or length.

An emerging area is text extraction from smartphone conversation (Ram et al., 2019)

Beyond allowing for a more nuanced assessment (Boyle & Hutchinson, 2009), text-based responses can provide higher measurement reliability than traditional closed-ended Likert scales (Jodoin, 2003; Kjell, Kjell, Garcia, & Sikström, 2018).

The obvious limitation: How to analyze text-based responses?

Example Response

Example response to "List 5 reasons for living"

"My innutity has not kicked in for my wife"
"My grand daughter needs a father image"
"Because I want to live to 161"
"Because my X wife is being supported by me and if I die that support goes away"
"Just because it is a good day."

Text Mining

Text mining could be summarized by detailing three main approaches:

1. Word or term frequency
2. Sentiment analysis
3. Topic modeling

The majority of research in psychology with text mining has focused on sentiment analysis or the use of dictionaries:
- This involves summarizing word use in comparison to dictionaries of emotion or valence categories.
- In psychology, this is Pennebaker's: Linguistic Inquiry and Word Count
- Also other standard dictionaries

Word Representation

Given a set of responses, the simplest way to represent the words is as a bag of words (BOW). This represents the words without respect to the order in which they were used. More complex approaches, often using deep learning, can represent the order, however, this requires larger sample sizes.

The second aspect: tree the words as singular units (unigrams) or as ordered pairs (bigrams). For the phrase:

– "I feel sad"

- unigram: "I", "feel", "sad"
- bigram: "I feel", "feel sad"

Dictionary Based

Dictionary based approaches involve defininining a dictionary of theoretically relevant words to the construct of interest. Kaitlin will discuss the use of Internal State Language and apply this dictionary to dialogues between mothers and children.

Pennebaker's LIWC is a dictionary based approach, using theoretically defined topic areas to describe text responses in a number of domains.

Once a dictionary is defined, this can easily be automatically applied in statistical software.

Sentiment Analysis

Sentiment analysis, is a method of text analysis to systematically identify and quantity the study affective state and subjective information in text.

Sentiment is mostly assessed at the document level through the use of one of my lexicons that assigns valence ratings to the individual words.

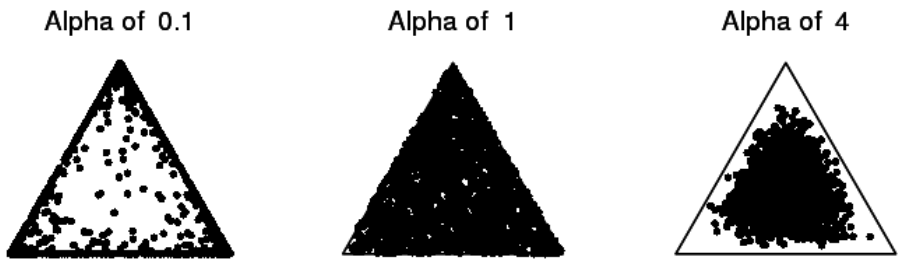
Aggregation across words provides a sentiment score. This score can then have utility either on its own, or by pairing with other aspects of the text.

Topic Modeling

Topic modeling extracts latent topics to summarize across documents or participant responses. This is similar to the use of mixture/LCA models.

The most common form: latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003).

This involves the use of the Dirichlet prior, determining the the uniqueness in the mixture of topics.

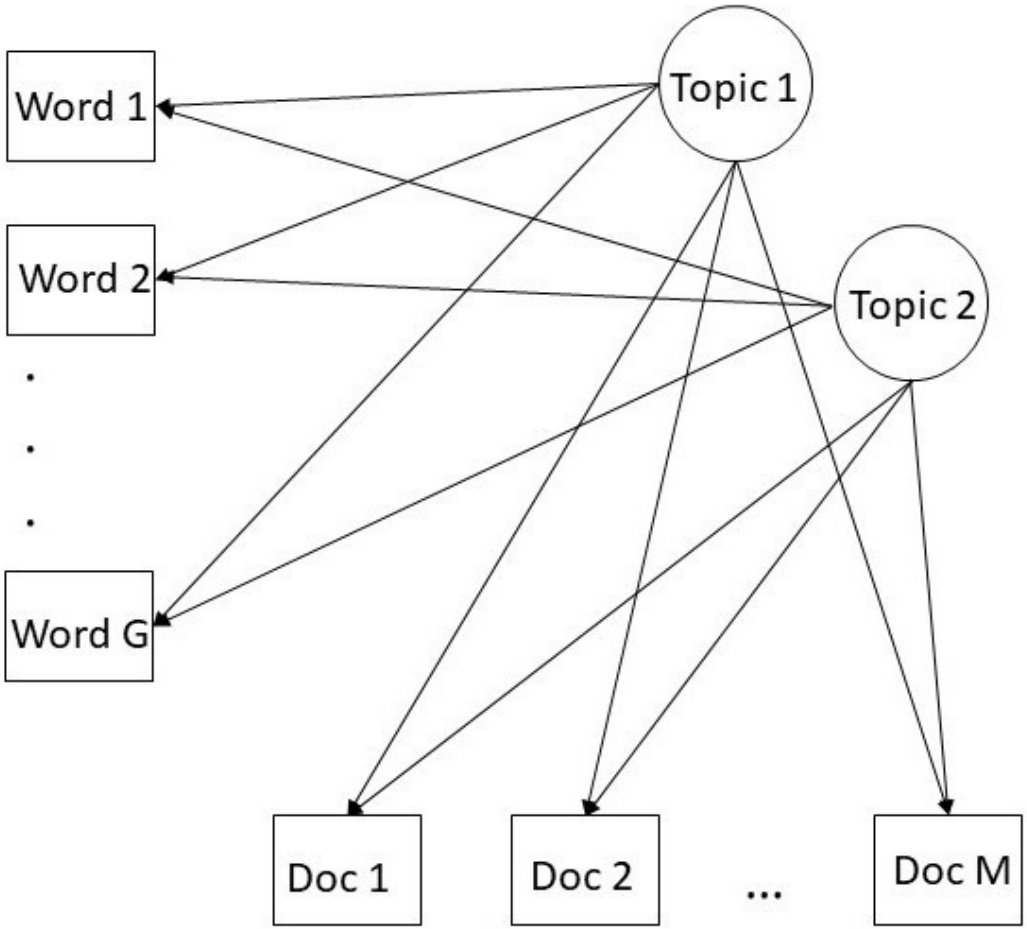


Topic Modeling

Instead of mathetically describing topic modeling, a high level summary:

- Choose the number of topics K
- Each word is a representation of K topics
- Each document is a representation of K topics

Topic Model Figure



LDA Flavors

LDA is most commonly performed as an unsupervised model.

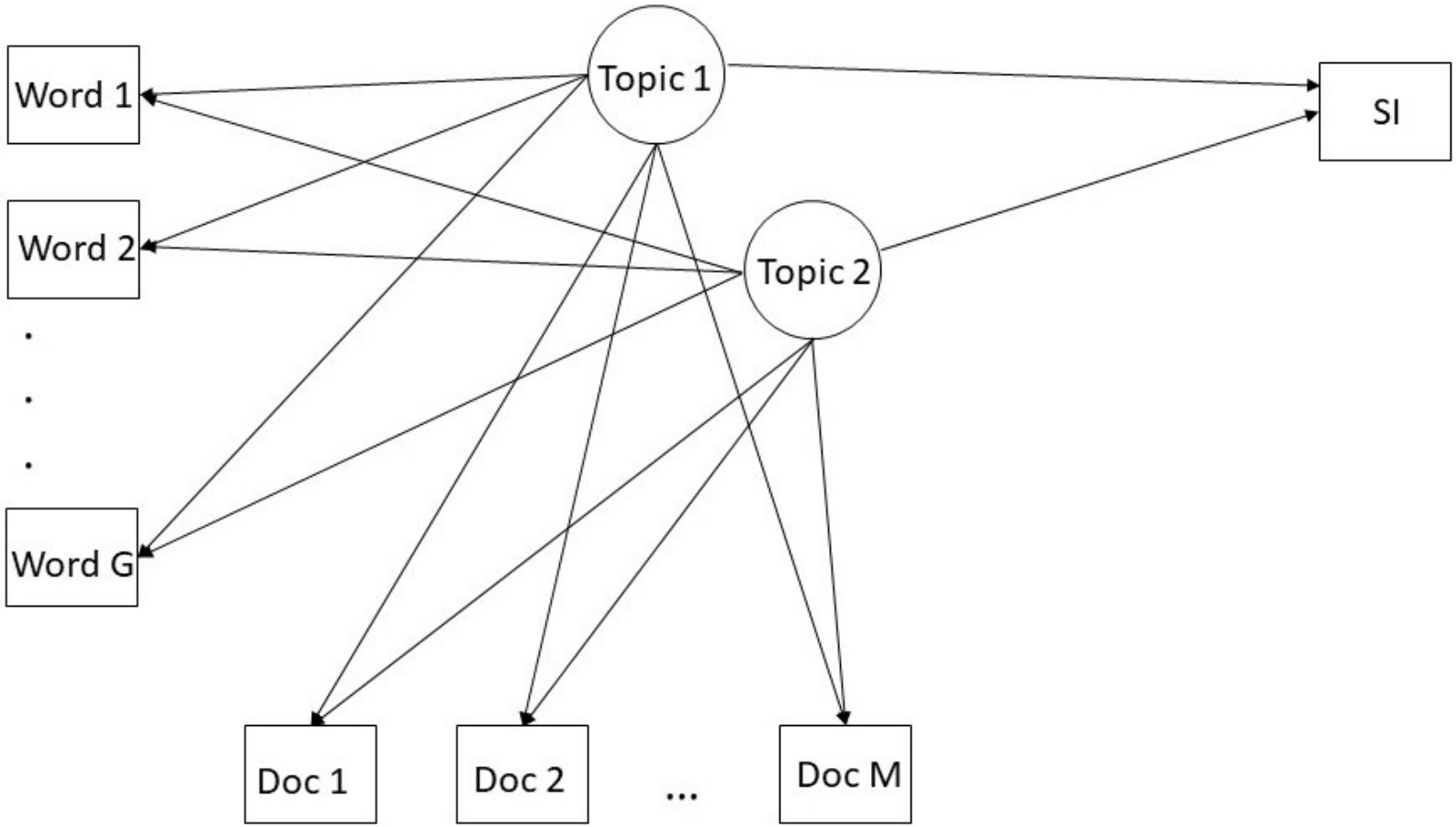
However, in a lot clinical research, we have an explicit outcome, or covariates of interest.

Other options include:
supervised LDA (sLDA; Blei & McAuliffe, 2008)
Integrating both into a joint model has shown superior predictive performance (Blei & McAuliffe, 2008; Zhu, Ahmed, & Xing, 2012)

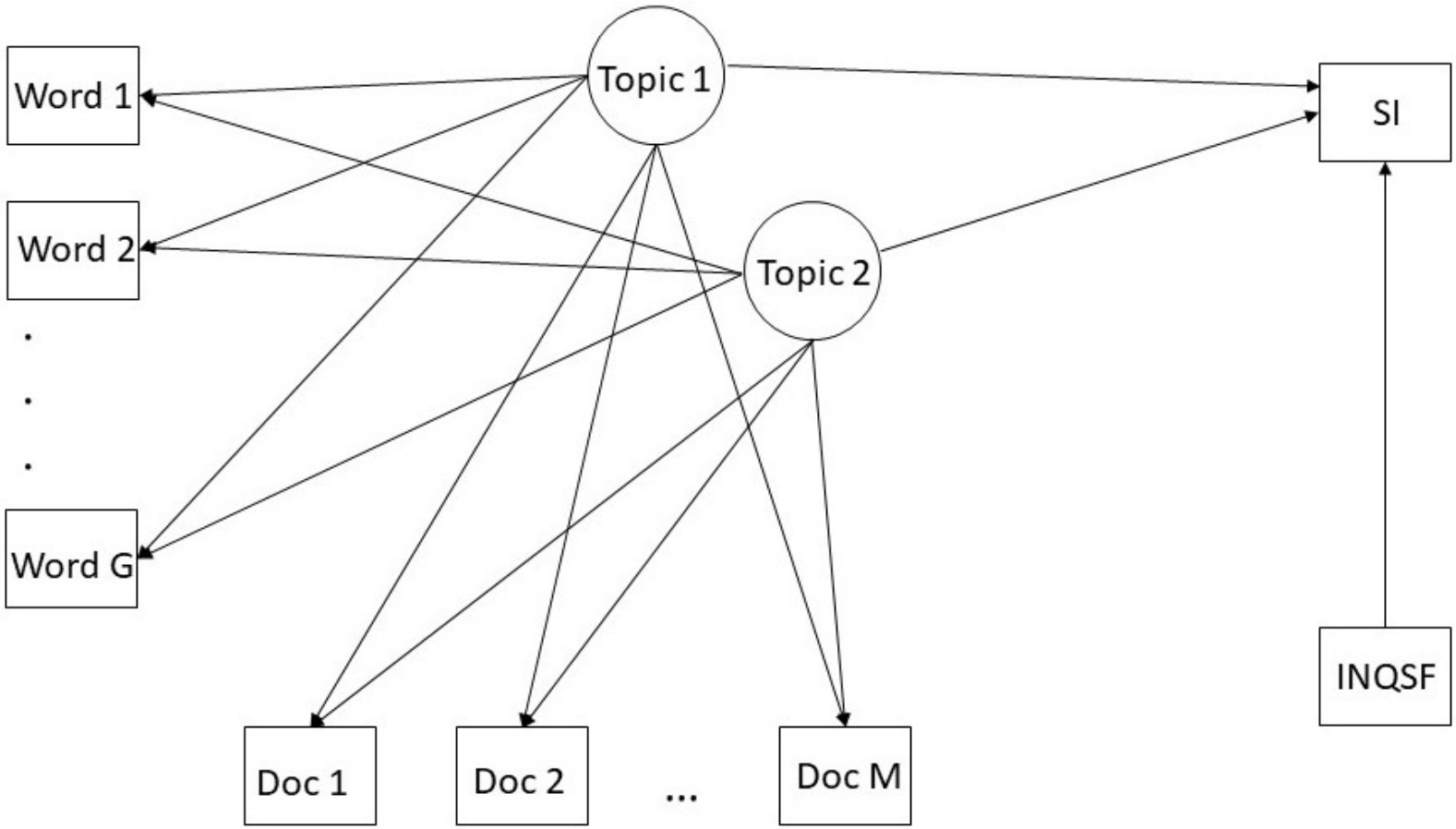
Adding covariates:
supervised LDA w/ covariates (sLDAX; Wilcox, Jacobucci, & Zhang, in prep)

Using covariates to explain the topics:
structural topic model (STM; Roberts et al., 2014)

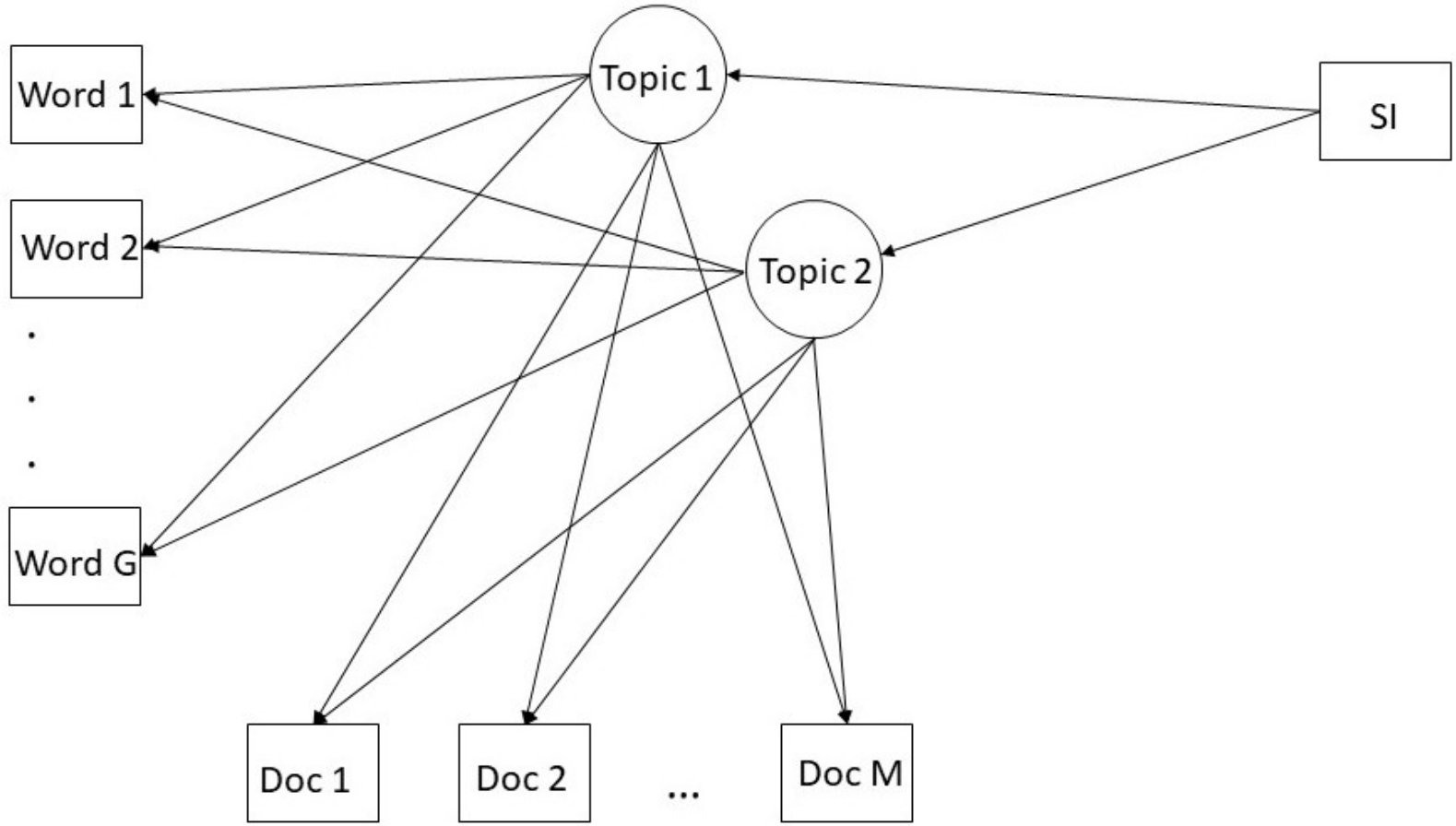
Supervised LDA



Supervised LDA w/ Covariates



Structural Topic Model



LDA Interpretation

Main focus: Top word representations for each topic
– similar to loadings in factor analysis

Can also assess what topics each document is most representative of
– Help classify observations according to topic.

Choosing the Number of Topics

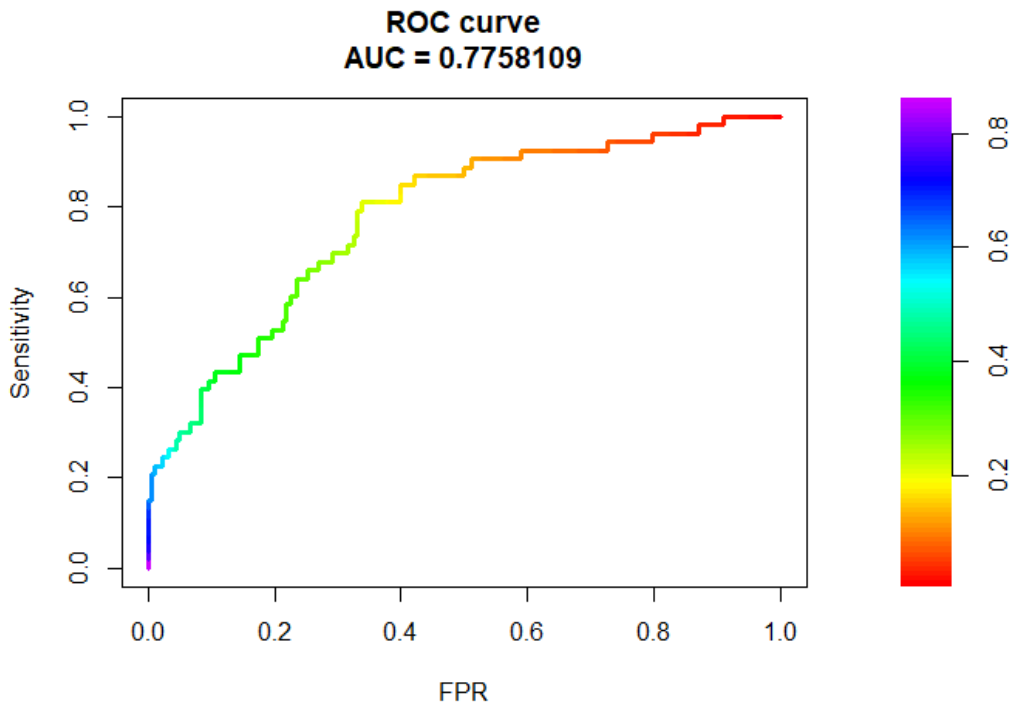
Good: Fewer metrics than in structural equation modeling
Bad: Fewer metrics than in structural equation modeling

Although metric exist, most notably perplexity, they more concerned with topic quality – We care more about explaining the outcome.

Therefore, we assess area under the receiver operating characteristic curve (AUC). With SLDAX, this is done with Bayesian estimation of the posterior.

Area Under the Curve

The area under the receiver operating characteristic curve (AUC) can be characterized as the probability that a randomly drawn positive case has a higher probability than a randomly drawn negative case (Fawcett, 2006).



Challenges:

- Missing data
- Small sample sizes
- How to ask quality questions
- Participant burden
- Determining # of topics
- Estimation algorithm
- Topics vs. sentiments vs. n-grams vs. phrases
- The use of stemming and stop-words
- Longitudinal data

Thank You!

Questions or material: rjacobuc@nd.edu
Slides posted: rjacobuc@nd.edu/presentations