A Step-by-Step Protocol for Curation of Topic Models by Subject Matter Experts

Philip Resnik[1], Pranav Goel[1], Alexander Hoyle[1], Rupak Sarkar[1], Josh Hagedorn[2], Maeve Gearing[3], Carol Bruce[3]

[1]University of Maryland, [2]TBD Solutions, [3]Westat

The Problem

- Content analysis is a dominant use case for topic models. In a random sample of 50 non-CS papers using topic models (2019-2022):
    - 94% : category discovery for human consumption
    - 68% : assign human-readable code labels
    - 64% : use topic probabilities as a form of coding

- Particularly for neural models, there is no valid automatic quality metric to optimize (Hoyle et al. 2021) and models can often lack stability. It is therefore difficult to trust that the automatic output of any particular model run will correspond to human judgment in a trustworthy way.
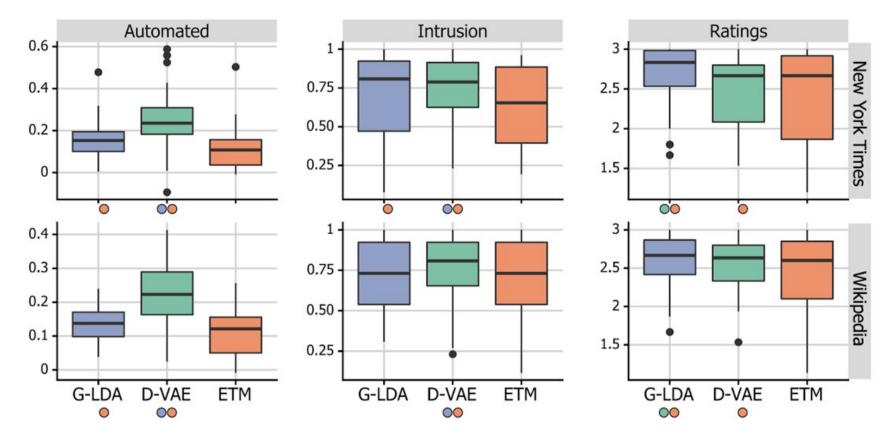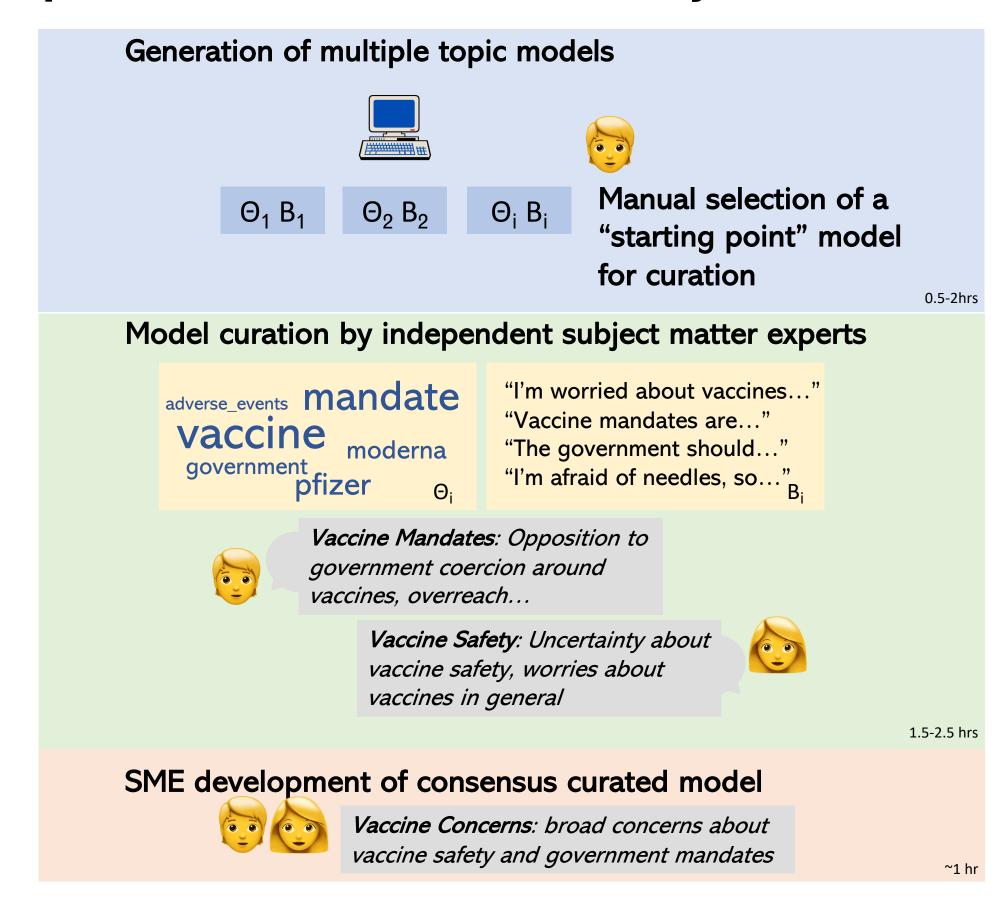

Figure 2: While automated evaluations (here, NPMI) suggest a clear winner between models, human evaluation is more nuanced. Human judgments exhibit greater variability over a smaller range of values. Colored circles correspond to pairwise one-tailed significance tests between model scores at $\alpha = 0.05$; for example, the rightmost orange circle at bottom right shows that human intrusion ratings for D-VAE are significantly higher than ETM for topics derived from Wikipedia.

- Qual practitioners lack a simple, consistent recipe for using topic models. This contributes further to a lack of trust.
    - Hoyle et al. 2021: standardization gap
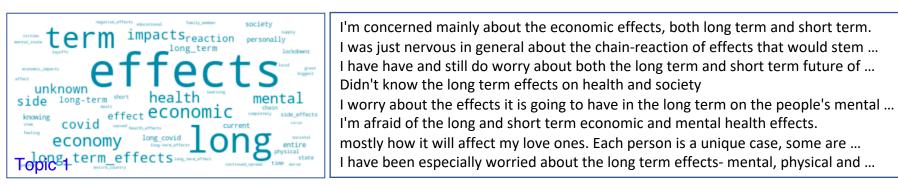   - Cf. Doogan 2022, de Volo et al. 2022

Computer Assisted Content Analysis Protocol

Generation of multiple topic models

$\Theta_1 B_1$ $\Theta_2 B_2$ $\Theta_i B_i$ Manual selection of a "starting point" model for curation

0.5-2hrs

Model curation by independent subject matter experts

adverse_events mandate vaccine moderna government pfizer Θ_i

"I'm worried about vaccines…"
"Vaccine mandates are…"
"The government should…"
"I'm afraid of needles, so…" B_i

Vaccine Mandates: Opposition to government coercion around vaccines, overreach…

Vaccine Safety: Uncertainty about vaccine safety, worries about vaccines in general

1.5-2.5 hrs

SME development of consensus curated model

Vaccine Concerns: broad concerns about vaccine safety and government mandates

~1 hr

The Protocol in Action

- Open-ended response from a U.S. nationally representative COVID-19 impact survey, N=1081: *Please tell us more about your concerns or worries about the COVID-19 pandemic, both in the early months of the pandemic and now.*
- Manual selection among models with K = 15, 20, 25, 30
- Curation of a 25-topic model

Topic	SME1 Label	SME1 Description	SME2 Label	SME2 Description
1	Societal effects	effects of the pandemic on economy, health and society, particularly in the long term	How do you worry about short and long-term impacts of COVID	Impact of COVID

Consensus Label	Consensus Description
Impact of COVID	Short and long-term effects on economy, health, and society (mostly negative)

Pilot Evaluation

		Jaccard Agreement by code merging criteria	
First coder	Second coder	Restrictive	Broad
Manual A	Manual B	0.385	0.720
Manual A	Consensus C	0.357	0.692
Manual B	Consensus C	0.348	0.652
Manual A	Assisted D	0.394	0.700
Manual B	Assisted D	0.444	0.654
Consensus C	Assisted D	0.387	0.612

Manual A, Manual B : results from two experienced SMEs conducting manual content analysis
Consensus C : results from our step-by-step protocol with two (other) experienced SMEs
Assisted D : freelance analyst doing the independent-SME assisted analysis

Current Take-Aways

- Considerable variability even in traditional analysis
- Strong (comparable?) performance using protocol…
- …With much less time taken.
- Looking at top *documents*, not just top words, is crucial

Limitations

- One English-only dataset with reasonably long responses
- Manual content analysts had access to entire response set
- Evaluation setting requires better specification

References

- Boyd-Graber et al. "Applications of Topic Modeling." Foundations and Trends in Information Retrieval. 2017.
- Doogan, C. A Topic is Not a Theme: Towards a Contextualized Approach to Topic Modelling. Doctoral dissertation, Monash University. 2022.
- Hoyle, Goel, et al. "Is automated topic model evaluation broken?". NeurIPS. 2021.
- Hu et al. "Interactive Topic Modeling". ACL. 2011.
- de Volo et al. "LDA in the Wild: How Practitioners Develop Topic Models." West Coast NLP. 2020
- Musialek, Chris, Philip Resnik, and S. Andrew Stavisky. "Using text analytic techniques to create efficiencies in analyzing qualitative data: A comparison between traditional content analysis and a topic modeling approach." Am. Assoc. for Public Opinion Research (2016).