“Why do you say that?” A Case Study in Applying Topic Modeling to Analyze Open-ended Survey Probes

Benjamin Rogers (He/Him), National Center for Health Statistics

Valerie Ryan (She/Her), National Center for Health Statistics

Kristen Cibelli Hibben (She/Her), National Center for Health Statistics

Advancing Efficiency and Accuracy with Data Science Techniques – 2022 FCSM Research & Policy Conference

"The findings and conclusions in this presentation are those of the author and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention"

Introduction

What is Topic Modeling?

- Unsupervised learning technique used to identify substantive themes that create the hidden semantic structure of a collection of documents
- Assuming each document has one or more defining and identifiable topic, but the topic is not labeled or provided for that document
- Topics can be found either by the distribution and frequency of words or observing the semantic similarity of the texts

Why use Topic Modeling?

- Manual exploration of text is time consuming
- Challenging to organize and identify substantive topics manually
- Provides an estimate for the prominence of different topics within the set of documents

Data

Research and Development Survey (RANDS)

- Ongoing series of surveys conducted by the NCHS Division of Research and Methodology (https://www.cdc.gov/nchs/rands/)
- Primarily recruited, web-based commercial survey panels
- 3 rounds of RANDS during COVID-19 were conducted during summer 2020 and spring 2021
- 12 open-ended web probes based on cognitive interview questions in RANDS during COVID-19 over the 3 rounds
- Total of 103,397 responses after filtering out Gibberish, Refusal, and Uncertain responses*

*"Toward a Semi-automated Item Nonresponse Detector Model for Open-response Data" - Hibben et al, Advancing Efficiency and Accuracy with Data Science Techniques , 2022 FCSM Research & Policy Conference

Pandemic Probe used in RANDS during COVID 19

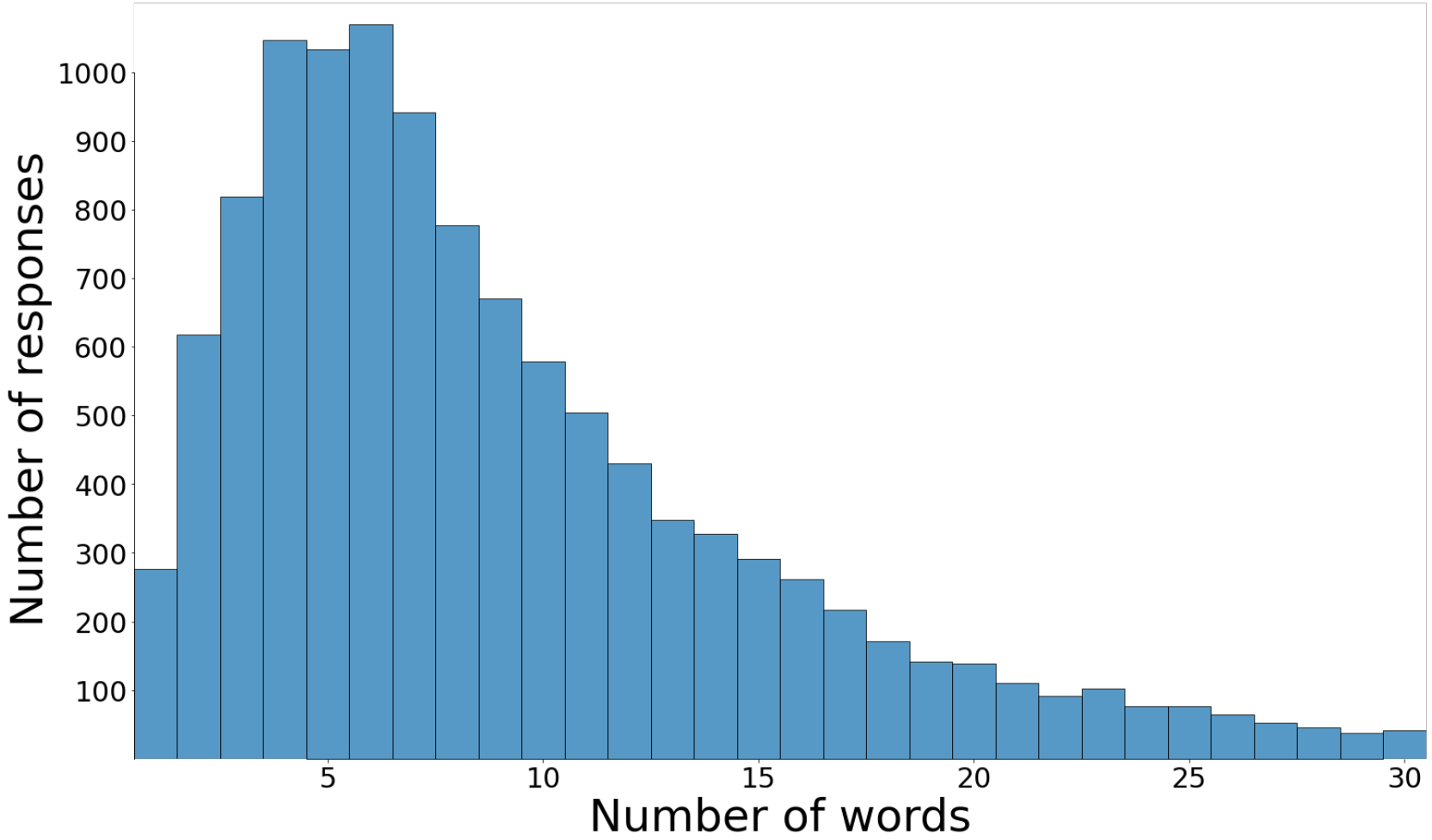
Round 1 – Summer 2020

- Question: “When do you think the Coronavirus pandemic began? Your best guess is fine”

- Question: “When did the Coronavirus pandemic first affect your daily life? Your best guess is fine.”

- **Question: “Why do you say that?”** <- We are studying responses to this probe

Majority of open-ended survey responses are short for the Pandemic Probe



- Median word count: 8
- Median word count for responses across all probes: 5
- Number of responses with more than 30 words: 143

Methods

Selecting Topic Modeling techniques to compare

- Many are currently used
  - Examples: Nonnegative Matrix Factorization, Probabilistic Latent Semantic Analysis, Hierarchical Dirichlet, Contextualized Topic Modeling, BERTopic
- Chose to compare LDA and Top2Vec
  - Two distinct underlying methodologies for identifying topics: similarity vs frequency
  - LDA represents traditional approach – frequency
  - Top2Vec represents recent advancements– similarity calculated from embeddings

Latent Dirichlet Allocation (LDA)*

- Generative probabilistic model using “bag-of-words” assumptions
  - Order of text is assumed to not matter
  - Does not integrate the meaning of words into model
  - Published in 2003
- Requires text preprocessing – stop word removal and lemming/stemming
  - Example: “that’s when we heard about the virus and started distancing”

    Output: “hear”, “virus”, “start”, “distance”
  - Example: “lock down and businesses closing”

    Output: “business", "close”

After preprocessing the median word count dropped from 8 to 4

*“Latent Dirichlet Allocation” – Blei et al 2003

Example output of LDA topics

First 7 topics – probability of the top 3 words of the topic

1. 0.207*people + 0.059* getting + 0.055*sick
2. **0.329*close + 0.206* school + 0.088*lockdown**
3. 0.241*place + 0.199* everything + 0.106*shut
4. 0.225*march + 0.087* become + 0.048*case
5. **0.235*social + 0.210* distance + 0.050*supply**
6. 0.081*city + 0.061* cause + 0.053*normal
7. **0.201*mask + 0.199* store + 0.111*grocery**

Top2Vec*

- Semantic similarity based – Assumes similarity of texts indicate topics
  - Uses document embeddings from pre-trained models to create high dimensional representations of the text
  - Can Identify the number of topics in a corpus automatically but can be provided number of topics
  - Requires no text preprocessing
  - Published in 2020

*“Top2Vec: Distributed Representations of Topics” Dimo Angelov (2020) [2008.09470] Top2Vec: Distributed Representations of Topics (arxiv.org)

Strengths and Weaknesses of the models

LDA

- Weaknesses
  - Does not handle misspelled words, acronyms, synonyms, or dates well
  - Performs poorly on short text
  - Topics identified can be difficult to decipher
- Strengths
  - Does not identify topics based on grammatical structure

Top2Vec

- Strengths
  - Handles dates, synonyms, acronyms, and misspelled words well
  - Identifies subtopics - “Gambling”
  - Identifies grammatical topics - “That is when ____”
- Weaknesses
  - Semantically similar does not always represent a substantive topic

Identifying number of topics to specify for modeling

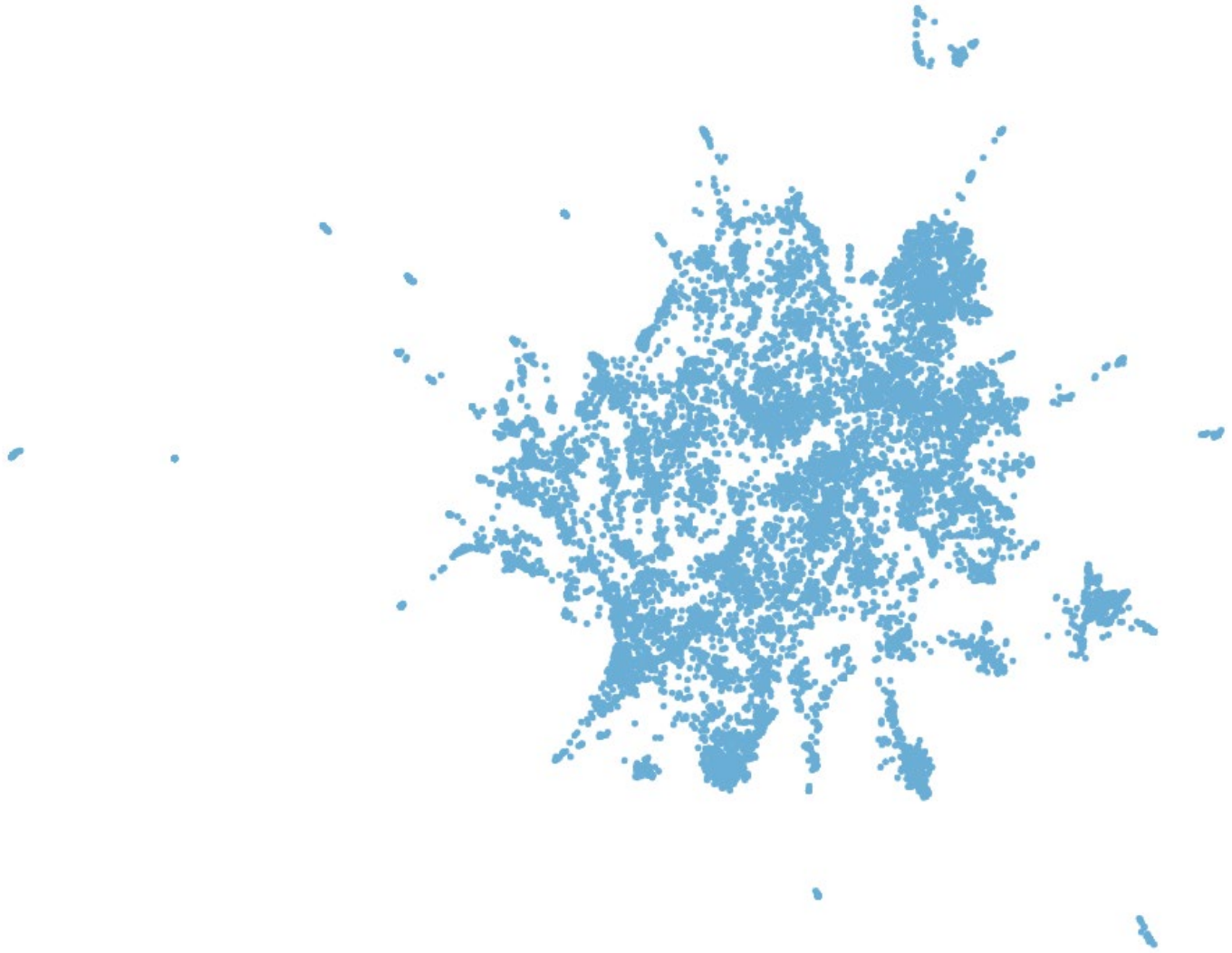
- No great option
  - Clustering the texts and counting clusters
  - Calculate coherence score for range of topic numbers and selecting “elbow point”
  - **Select a high number of topics and then have subject matter experts review for number of broad topics**
  - Have a predetermined number by subject matter experts before modeling

After review of a Top2Vec model with 148 topics, subject matter expert review identified 26 topics

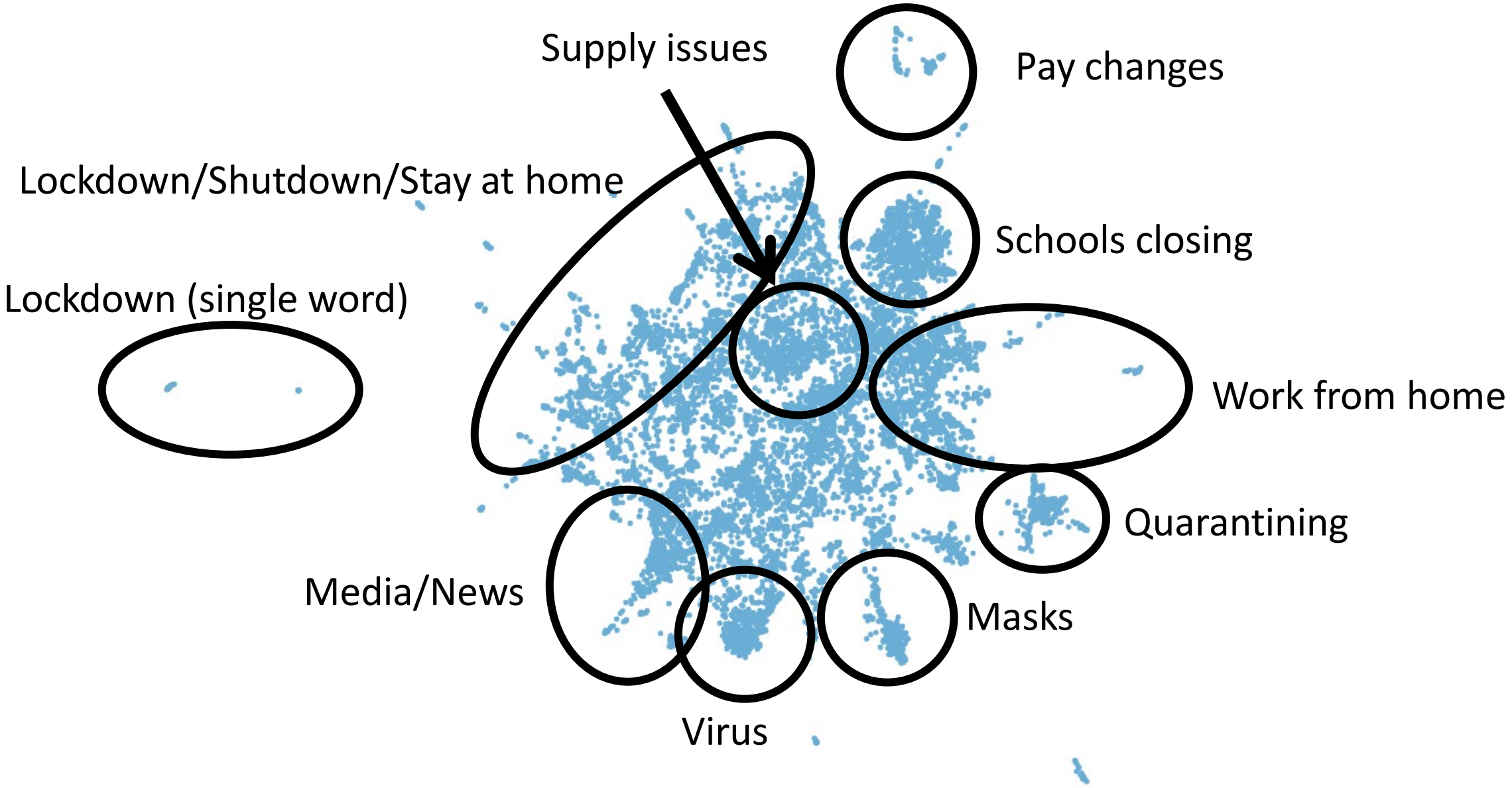
- Difference between semantically similar versus substantive topics

Visualization of Pandemic Responses

UMAP visualization of the Pandemic Probe responses



Topic clusters present in the Pandemic Probe



Top2Vec Topics

LDA Topics

Clusters of colors illustrate distinct topics identified by the model

Topics Found

Grammatical Structure vs Similar Meaning

“That’s When”/“Thats when”/“That is when”

- 1,921 responses starting with these variations
- LDA removes with preprocessing
- Top2Vec identifies as similar – created grammatical topics

“Lock down” / “Lockdown”

- 113 responses with these variations
- LDA cannot recognize their similarity – different topics
- Top2Vec understands the semantic similarity – same topic

Example topics found from topic modeling

Topic	Example Response
Employment/Job	“laid off from job”
Lockdown/Shutdown	“everything shut down”
Schools closed	“Child’s school closed”
Masks	“Had to start wearing masks around”
Media	“Cases started being reported on news”
Social distancing	“Started needing to distance in public”
Birthday/Events	“Was around my birthday”
Canceled trips	“Trip to Europe was canceled then”

How should we use Topic Modeling with survey response?

- Does not replace human coding of responses
- Powerful tool with exploration/discovery of topics to use for human coding
- Top2Vec enables you to look at how people are answering questions
  - People respond with events they remember, changes to daily routine, or when COVID directly impacted their lives
  - Provides substantive and semantic topics

Manual selection of substantive topics from those generated is still required

Acknowledgements to Valerie Ryan and Kristen Cibelli Hibben for their assistance with this project

Questions?

Name: Benjamin Rogers

Email: qtw4@cdc.gov