CS 293/EDUC 473

Discovery & Exploration in Educational Text Data

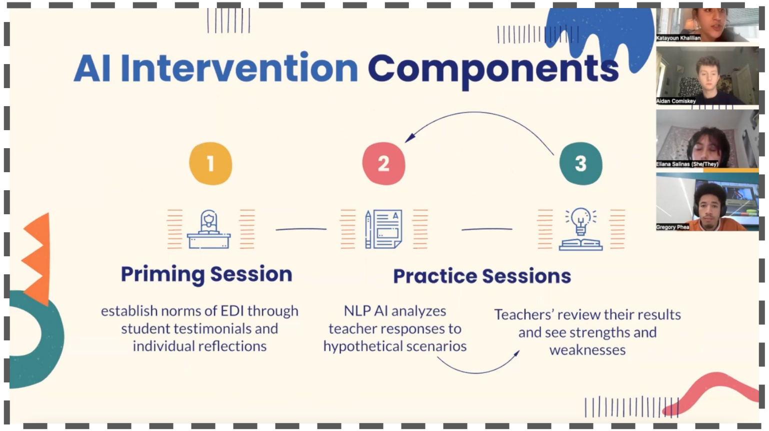
Topic Modeling, Clustering, Grounded Exploration

Stanford
GRADUATE SCHOOL OF
EDUCATION

Stanford | NLP

Round #1 practice pitches next Monday

- **4 mins per team**
- **Rubric**
- Example video (thanks to David Yeager & students) shared on Canvas



AI Intervention Components

1 Priming Session

establish norms of EDI through student testimonials and individual reflections

Practice Sessions

2 NLP AI analyzes teacher responses to hypothetical scenarios

3 Teachers' review their results and see strengths and weaknesses

Instructions for Practice Pitch & Peer Reviews

1. Add your slides to this deck (we'll tell you which session [Mei's or Dora's] you'll be assigned to)
2. To complete peer reviews:
   a. Go to Canvas > Practice pitch round 1 assignment
   b. You should see an option to give feedback on a peer's assignment
   c. Please fill out the rubric: 3 items, total 12 points
   d. In addition, please type in all feedback for each item OR upload pdf of this filled out rubric to Canvas.

What are you most excited to learn about?

Practical & Ethical Considerations

Generative AI & Education

Teacher Empowerment

Measurement

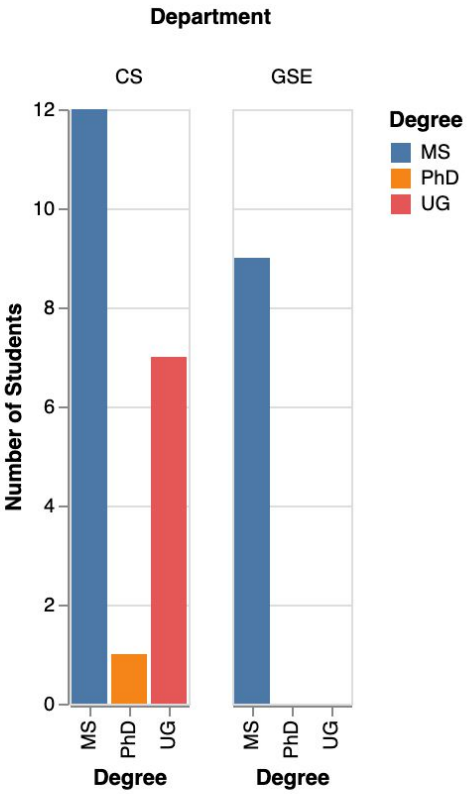
Classroom Impact

Guest speakers & Industry insights

Intersection of NLP, HCI & Ed tech

Project-based Learning

What degree program are you all pursuing?



Today’s class

- Brief activity around teacher interview
- Debrief two most recent guest visits by Dan Meyer & Google Learn LM team
- Brief lecture on topic modeling & clustering
- Reading discussion for Kubsch et al (2023) led by Ari, Eban, Khaulat and Ziqi

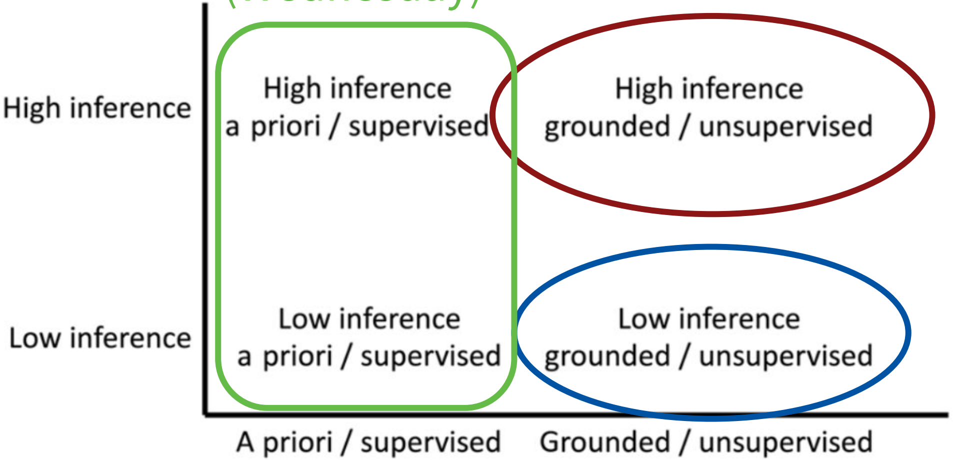
Group discussion on teacher interviews!

- **[Context]** What did you learn about the day-to-day realities of teaching that you hadn’t considered before?
- **[Challenges]** What challenges or pain points did the educator describe?
- **[Opportunities]** How do these challenges relate to potential opportunities for NLP tools or interventions?
- **[Human-Centered Design]** How did the educator describe their experiences with technology in the classroom, and what did you learn about the design of tools for education?

Talk in small groups

- Did anything surprise you, or challenge your ideas based on the readings & talks?
- Is there anything you disagree with?

Topic Modeling, Clustering, Grounded Exploration



classification (Wednesday)

topic modeling/thematic analyses

lexical analyses

	A priori / supervised	Grounded / unsupervised
High inference	High inference a priori / supervised	High inference grounded / unsupervised
Low inference	Low inference a priori / supervised	Low inference grounded / unsupervised

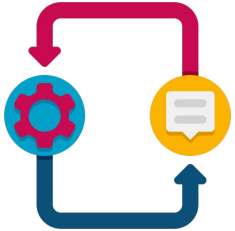
FIGURE 2 Epistemic functions and level of inference of four different kinds of tasks in assigning codes or numbers to data based on characteristics of the data and question

Kubsch et al. (2023)

When to use topic modeling & clustering?

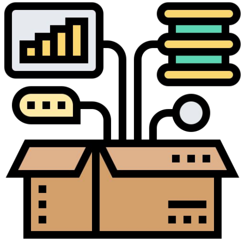
Discovery

Discovering **interesting or unexpected structures** that can be useful for hypothesis generation.



Synthesis

Comparing and synthesizing **qualitative** and **quantitative** analyses.



Featurization

Using topics as **features** (representations) in downstream analyses, e.g. log odds, LLM-based analyses, classification.

Your take from HW1:
Topics/clusters are very hard to interpret!

From HW1: Cons of **LDA**

- word co-occurrence may not always equal thematic consistency
  - math topics “blend together”
  - classroom management + math-related themes co-occurring
  - function words, transcription issues add a lot of noise
- difficult to interpret topics
  - words are removed from context
  - need domain knowledge
- hard to steer
  - not clear how to identify right number of topics
  - processing data in a way that leads to cleaner topics is nontrivial

From HW1: Cons of **clustering**

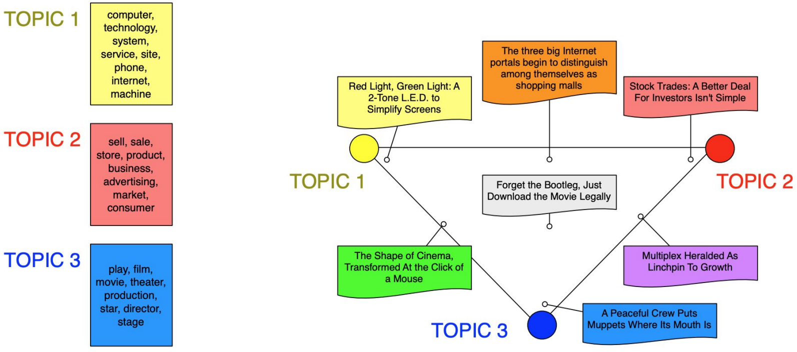
- very sensitive to surface-level lexical similarity
- very sensitive to length
- clusters become too fine-grained quickly (long tail…)

Our response (and some of yours, too…)

- processing the data and choosing the right model can make a big difference
    - models you tried in HW1 may not be the best
    - subsetting the data, removing irrelevant words can help improve them
    - pay attention to document length!
- trade-off between output interpretability (LLMs win here) vs model interpretability (LDA, clustering win here)
- sometimes you *surface level similarity* may be what you want to capture (example later)

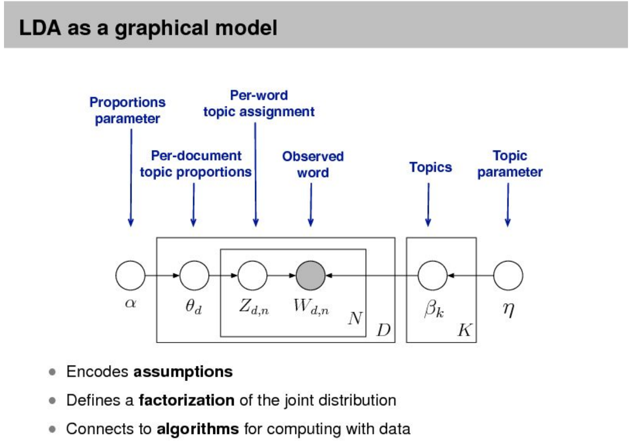
Topic modeling

- There are several topic modeling approaches; in my experience, **LDA MALLET** (David Blei) works best for simple explorations
    - Original command line package
    - Python wrapper
- LDA outputs per-document topic proportions and per-topic word proportions
- Take a look at these slides here to learn more



TOPIC 1
computer, technology, system, service, site, phone, internet, machine
TOPIC 2
sell, sale, store, product, business, advertising, market, consumer
TOPIC 3
play, film, movie, theater, production, star, director, stage
Red Light, Green Light: A 2-Tone L.E.D. to Simplify Screens
The three big Internet portals begin to distinguish among themselves as shopping malls
Stock Trades: A Better Deal For Investors Isn't Simple
TOPIC 1
Forget the Bootleg, Just Download the Movie Legally
TOPIC 2
The Shape of Cinema, Transformed At the Click of a Mouse
Multiplex Heralded As Linchpin To Growth
TOPIC 3
A Peaceful Crew Puts Muppets Where Its Mouth Is

(a) Topics (b) Document Assignments to Topics



LDA as a graphical model

- Encodes **assumptions**
- Defines a **factorization** of the joint distribution
- Connects to **algorithms** for computing with data

Credit: Jason Yosinski

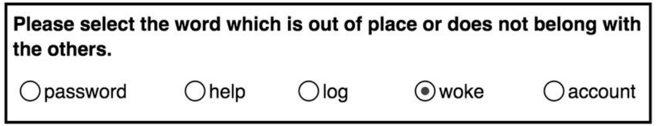
Pre-processing decisions for topic modeling **matter**

- Highest probability words within topics are often stop words (e.g. “I”, “not”, “will”) → **remove stopwords (?)**
    - Sometimes stopwords lead to meaningful distinctions (negations, use of 1st vs 3rd person)
- Forms of the same word may appear together as highest probability words (e.g. “like”, “liked”, “liking”) → **stem / lemmatize words (?)**
    - but: they may decrease stability (Schoefield & Mimno, 2016) and can lead to combining words with different meanings
- **Lowercasing?**
    - may collapse named entities with non-named entities; removes emphasis (e.g. all caps)
- (Denny & Spirling, 2017) introduce an R package **preText** to measure the impact of preprocessing on topics

the
a
of
in
and
to
my
on
rain
with
i
moon
her
from
for
day
night
snow
winter
summer
morning
spring
autumn
light
wind
water
sky
old
no
at
all
is
we
through

Evaluating topic models: Reading tea leaves (Chang et al., 2019)

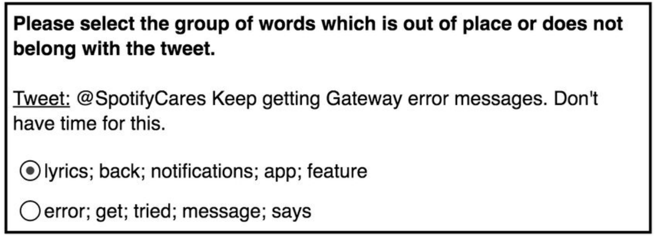
- Manual methods are the most reliable!
- Two evaluation tasks
  - **Word Intrusion**
  - **Topic Intrusion**
    - More challenging with longer texts
- Relatively quick once it's set up
  - For resource-efficiency, first select ~3 most promising parameter settings (num topics + preprocessing decisions) by eyeballing + automated metrics and compare those



Please select the word which is out of place or does not belong with the others.

○ password ○ help ○ log ● woke ○ account

Word Intrusion Task



Please select the group of words which is out of place or does not belong with the tweet.

Tweet: @SpotifyCares Keep getting Gateway error messages. Don't have time for this.

● lyrics; back; notifications; app; feature

○ error; get; tried; message; says

Topic Intrusion Task

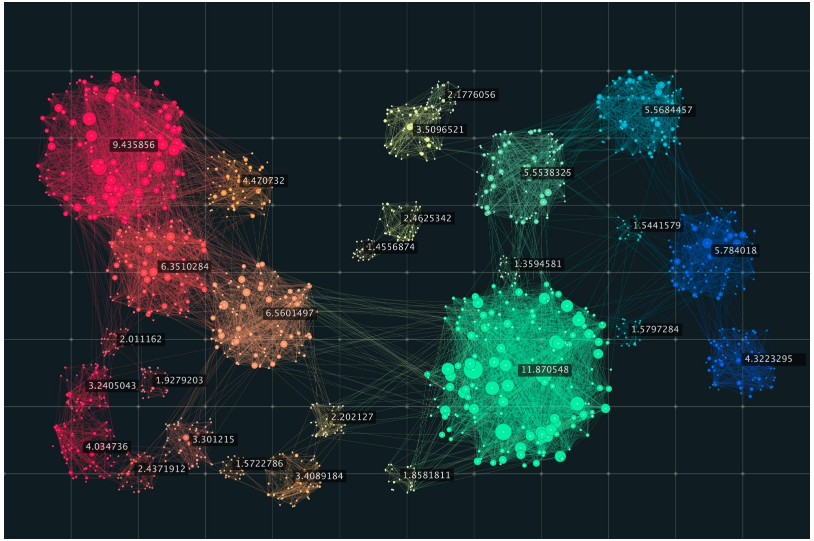
image credit: (Guzman et al., 2017)

Clustering

Group a set of data points into a number of clusters, so that

- Data points in the same cluster are similar to each other
- Data points in different clusters are dissimilar

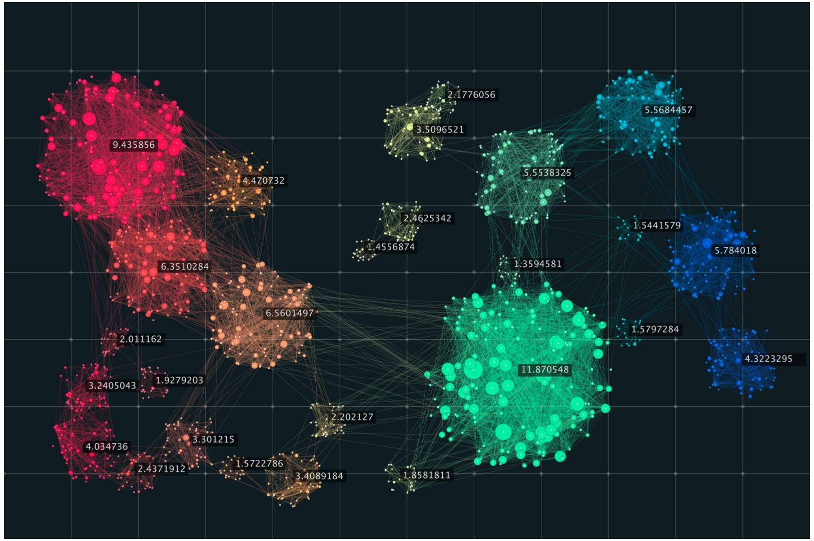
Use only X, not Y → unsupervised method



Slide credit: Diyi Yang

Cluster structures

- Partitioning a group of data point into K disjoint sets (K-means clustering)
- Assigning X to hierarchical structures (Hierarchical clustering)
- Assigning X to partial membership in K different sets (Graphic models, GMM)
- Learning a representation that puts similar data points closer to each other (Word embeddings w/ deep learning)

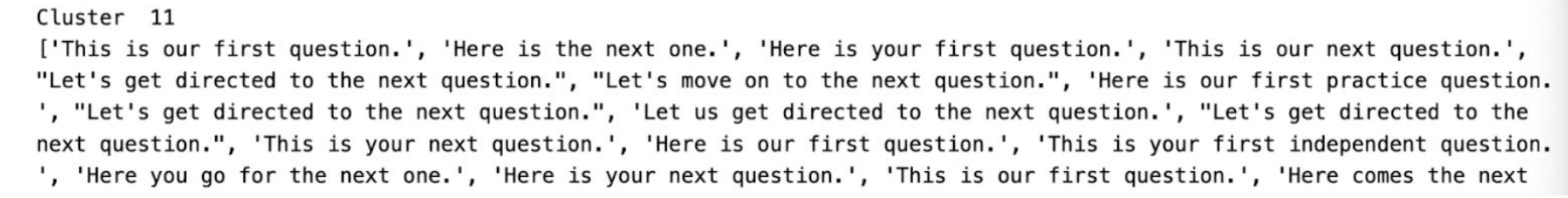


Slide credit: David Bamman & Diyi Yang

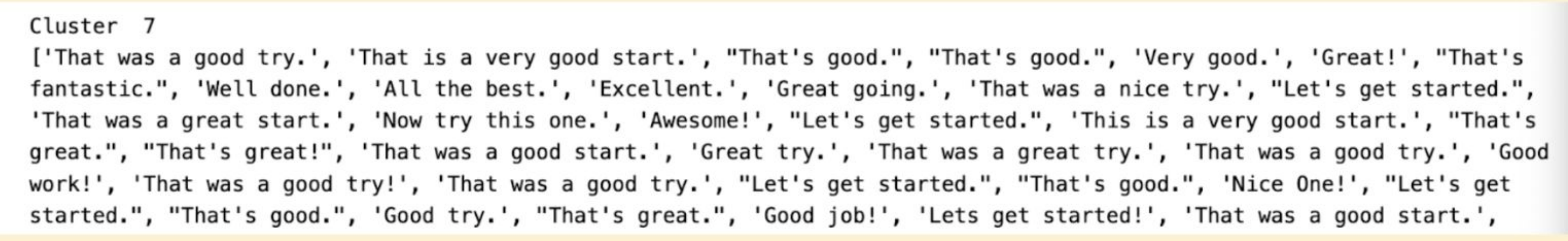
Word embedding based representations

- Simple but tough to beat baseline sentence embeddings
  - tf-idf weighted average of word embeddings
- Sentence transformers (see HW1)

Clustered tutor responses (Wang et al., 2023)



```
Cluster  11
['This is our first question.', 'Here is the next one.', 'Here is your first question.', 'This is our next question.',
"Let's get directed to the next question.", "Let's move on to the next question.", 'Here is our first practice question.
', "Let's get directed to the next question.", 'Let us get directed to the next question.', "Let's get directed to the
next question.", 'This is your next question.', 'Here is our first question.', 'This is your first independent question.
', 'Here you go for the next one.', 'Here is your next question.', 'This is our first question.', 'Here comes the next
```



```
Cluster  7
['That was a good try.', 'That is a very good start.', "That's good.", "That's good.", 'Very good.', 'Great!', "That's
fantastic.", 'Well done.', 'All the best.', 'Excellent.', 'Great going.', 'That was a nice try.', "Let's get started.",
'That was a great start.', 'Now try this one.', 'Awesome!', "Let's get started.", 'This is a very good start.', "That's
great.", "That's great!", 'That was a good start.', 'Great try.', 'That was a great try.', 'That was a good try.', 'Good
work!', 'That was a good try!', 'That was a good try.', "Let's get started.", "That's good.", 'Nice One!', "Let's get
started.", "That's good.", 'Good try.', "That's great.", 'Good job!', 'Lets get started!', 'That was a good start.',
```

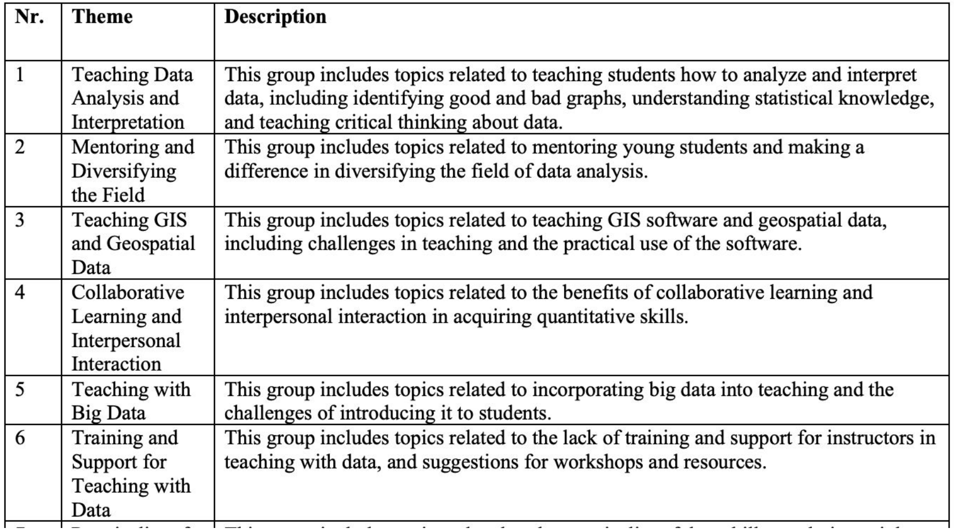
What to think about when doing clustering?

- How to **represent** each data point?
- How to **calculate the similarity** between data points?
- What is the **number of clusters** to use?
- How can we **evaluate** the resulting clusters?

Slide credit: Diyi Yang

Large Language Models for thematic analyses

E.g. Thematic Analysis (De Paoli, 2023; Gamieldien et al., 2023), Topic Modeling (Wang et al., 2023)



Nr.	Theme	Description
1	Teaching Data Analysis and Interpretation	This group includes topics related to teaching students how to analyze and interpret data, including identifying good and bad graphs, understanding statistical knowledge, and teaching critical thinking about data.
2	Mentoring and Diversifying the Field	This group includes topics related to mentoring young students and making a difference in diversifying the field of data analysis.
3	Teaching GIS and Geospatial Data	This group includes topics related to teaching GIS software and geospatial data, including challenges in teaching and the practical use of the software.
4	Collaborative Learning and Interpersonal Interaction	This group includes topics related to the benefits of collaborative learning and interpersonal interaction in acquiring quantitative skills.
5	Teaching with Big Data	This group includes topics related to incorporating big data into teaching and the challenges of introducing it to students.
6	Training and Support for Teaching with Data	This group includes topics related to the lack of training and support for instructors in teaching with data, and suggestions for workshops and resources.

Pros:

- Can do very well at producing highly interpretable topics with good coverage

Downside:

- Hallucinations
- Black box process; the outputs require extensive evaluation to ensure they're accurately representing the data

Topic modeling can support downstream supervised learning

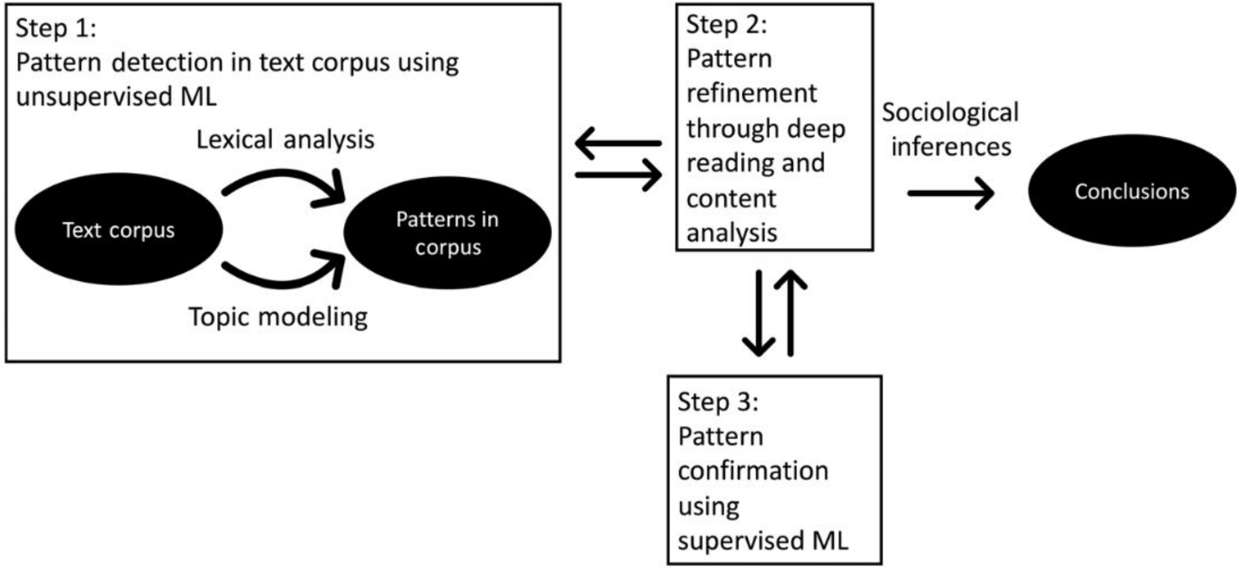

Step 1:
Pattern detection in text corpus using unsupervised ML
Lexical analysis
Text corpus
Patterns in corpus
Topic modeling
Step 2:
Pattern refinement through deep reading and content analysis
Sociological inferences
Conclusions
Step 3:
Pattern confirmation using supervised ML

FIGURE 4 Analysis workflow from Nelson, 2020

Reading Discussion for Kubsch et al. (2023)