TEXT IS ALWAYS BIASED

Julian Barg

Overview

1. Parallels between NLP & methods debate

2. Case study

3. Standard topic modeling/NLP approach

4. The challenge

5. Wrangling the data

6. Implications

7. Outlook

Natural Language Processing & methods debate — parallels

A story on two levels

- Micro vs. macro
  - Need to keep the big picture in mind
  - But dominant big picture narratives may be inaccurate
  - ➡️ Critical judgement required
- NLP
  - Ambition: process big amounts of data
  - Output is the dominant view, not the *accurate* view
  - ➡️ Critical judgement required

Example: "rotten" Corporate climate debate

- Applied example
  - Think of a company with a poor environmental/social impact

Ivey

Natural Language Processing & methods debate — parallels

A story on two levels

- Micro vs. macro
  - Need to keep the big picture in mind
  - But dominant big picture narratives may be inaccurate
  - ➡️ Critical judgement required
- NLP
  - Ambition: process big amounts of data
  - Output is the dominant view, not the *accurate* view
  - ➡️ Critical judgement required

Example: "rotten" Corporate climate debate

- Applied example
  - Think of a company with a poor environmental/social impact
- All corporations have climate action policy
- Are very effective at getting message out, dominate discourse
- Macro: we have the receipt — record global emissions
  - ❗ There is a problem, but what is its root?
- Micro: have to form our own judgement

➡️ Opportunity for org scholars to be at the forefront

Ivey

Case study: Keystone XL

Case study: deviating perception of a "climate victory" across micro and macro level

- National discourse:
  - "*The deepest thanks, however, go to you: to indigenous peoples who began the fight, to the folks in Nebraska who rallied so fiercely, to the scientists who explained the stakes [and five more groups].*" (McKibben, 2011)

- Local discourse:
  - Senator C.: "*But they influenced these people who love the land to join with them and join with them with those who are against all of these things, those that are against pipelines.*"

Ivey

Topic modeling standard approach — picture a sieve …

Topic modeling as linear process with some iteration

Take text documents, boil down to information

1. Select documents, *render* into structured *corpus*

2. Use topic modeling to *render topics*

3. Use topics to study corpus (*render theoretical artifacts*)

(Hannigan, Pak, and Jennings, 2022)

Ivey

… but institutional research should snowball

Topic modeling as linear process with some iteration

- ❗ What if we do not know the context well enough? Artifacts will have less nuance than original data.
- ❗ We can set the level of detail of output, but topic model controls contents.
- How to retain niche counter narratives when the official narrative is off?

Information that would not have made it

- Individuals being snubbed for their positions
  - Chairman L.: "*Good. Are there any questions for Mr. Swift? Seeing none, thank you very much. Well done.*"
- Topics being pushed aside
  - Mr. W.: "*The Nebraska Chapter of 350.org … is very concerned about global warming and climate change …*"
  - Senator L.: "*… you brought up a whole new concept I hadn't thought of. Is there any AG review of the committee?*"
- Struggles over identity and legitimacy
  - Mrs. L.: "*I would like to note that I'm not a radical and I'm not an extremist as suggested of some of us by some of those testifying on Monday*"
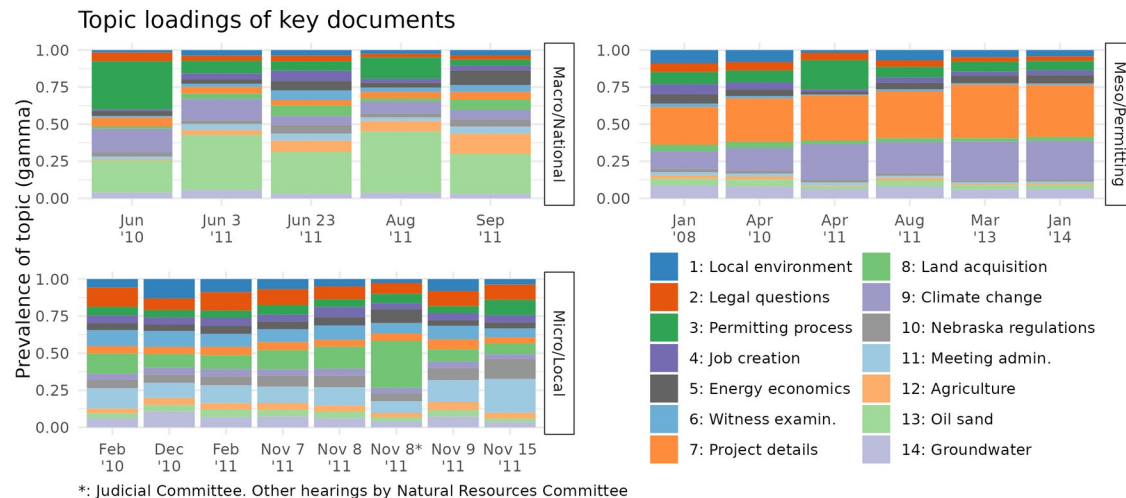
Ivey

Examples of a mixed-methods workflow

- Update topic model in response to qualitative findings
    - Correct topic models for bias (missing topics)
    - Render corpus with out-of-sample documents

- Use topic loadings to aid qualitative work
    - Scripts: generate arbitrary topic loadings
    - Dashboards: inspect loadings and data side by side

Illustration: Topics across levels, generated in qual stage

Ivey

Examples of a mixed-methods workflow

- Update topic model in response to qualitative findings
  - Correct topic models for bias (missing topics)
  - Render corpus with out-of-sample documents
- Use topic loadings to aid qualitative work
  - Scripts: generate arbitrary topic loadings
  - Dashboards: inspect loadings and data side by side

Illustration: Topics across levels, generated in qual stage

Topic loadings of key documents

Prevalence of topic (gamma)

Macro/National — Jun '10, Jun 3 '11, Jun 23 '11, Aug '11, Sep '11

Meso/Permitting — Jan '08, Apr '10, Apr '11, Aug '11, Mar '13, Jan '14

Micro/Local — Feb '10, Dec '10, Feb '11, Nov 7 '11, Nov 8 '11, Nov 8* '11, Nov 9 '11, Nov 15 '11

*: Judicial Committee. Other hearings by Natural Resources Committee

1: Local environment
2: Legal questions
3: Permitting process
4: Job creation
5: Energy economics
6: Witness examin.
7: Project details
8: Land acquisition
9: Climate change
10: Nebraska regulations
11: Meeting admin.
12: Agriculture
13: Oil sand
14: Groundwater

Ivey

Implications

- Specific workflow may be impractical for many empirical studies
  - Promised efficiency of NLP does not materialize if we still conduct qualitative analysis
- But we need to attend to the general problem

- ❗ in 2–3 years, we will see use of ChatGPT in qualitative work in publications
  - ➡️ How do we avoid that we just regurgitate talking points?
  - ➡️ How to maintain nuance?
- ❗ How do we apply critical judgement to that workflow?
- ❗ How do we prevent an avalanche of studies that just reproduce dominant narratives?

Ivey

Some simple fixes

Back to the story on two levels

- Micro vs. macro
  - Find meaning at macro level
  - Identify corresponding data at micro level
  - Be prepared to do it over when micro and macro are inconsistent
- NLP
  - Be intentional
  - Don't let data dictate questions, but interrogate data with questions
- ➡ Critical judgement required

Example: "rotten" Corporate climate debate

- For ChatGPT & co: improved prompt engineering
  - Provide information on pitfalls when using NLP tools
  - Let's return to the example from earlier

Ivey

References

Hannigan, T. R., Pak, Y., and Jennings, P. D. (2022). Mapping the Multiverse: A Cultural Cartographic Approach to Realizing Entrepreneurial Possibilities. *Research in the Sociology of Organizations*, 217–237.

McKibben, B. (2011). Um, we won. You won. https://350.org/breaking-news-keystone-rejected-we-won-you-won-thank-you/.