Introduction to Computational Text Analysis (for Social Science)

Miriam Hurtado Bodell

LINKÖPING UNIVERSITY

“Big data” → computational social science

Massive increase of (digital) unstructured text data, for example;

- New social infrastructures → more aspects of social life happens online
- Digitalization efforts by for example national libraries and archives

Combined with extremely rapid methodological developments (machine learning + natural language processed) =
new possibilities for social scientific inquiry!

"Big data" → computational social science

Massive increase of (digital) unstructured text data, for example;

- New social infrastructures → more aspects of social life happens online
- Digitalization efforts by for example national libraries and archives

Combined with extremely rapid methodological developments (machine learning + natural language processed) =
new possibilities for social scientific inquiry!

BUT will never replace the need to read yourself.

"Big data" → computational social science

This course is about methods, but they are only part of the story:

- Identify a relevant and interesting research question
- Decide on research design (what would you do if you had infinite resources vs. what can you do)
- Select data (what is it a sample of?)
- Select method to answer RQ

Or, use a method to explore data to identify a research question (inductive research)

Why use text as data in social inquiry?

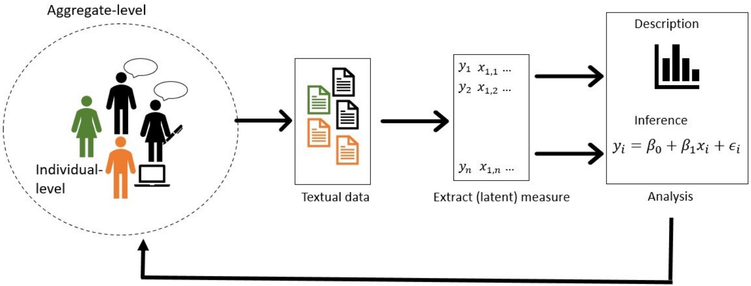
Social science wants to document and explain aspects of society and relations between people → text by itself is typically not useful

Text as a social sensor is useful for social scientists

- Aggregate-level; public culture (e.g. newspapers, social media)
- Individual-level; private culture (e.g. social media posts, diaries)

Why use text as data in social inquiry?

We are generally interested in latent properties of society or individuals



Choosing data

Choosing data depends on the goal of your research; what/who do you want to describe or make inferences about, e.g.

- Who is active on the online forum?
- What does the front page of a newspaper represent?
- What is a text unit?

Text data can be erroneous in different ways (especially text that has been digitized)! Make sure that the quality is good enough to answer your RQ

Three main tasks (Grimmer et al. 2023)

Helpful to divide different methods for computational text analysis by task instead of tool = focus on what we want to do instead of the latest method (a new method will exist before your paper is published)

- **Discovery**; let the data speak
- **Measurement**; extract (latent) features
- (Causal) **Inference**; explain or predict patterns
- *Generate or simulate; create new texts data or simulate conversations*

Three main tasks (Grimmer et al. 2023)

Helpful to divide different methods for computational text analysis by task instead of tool = focus on what we want to do instead of the latest method (a new method will exist before your paper is published)

- **Discovery**; let the data speak
- **Measurement**; extract (latent) features
- (Causal) **Inference**; explain or predict patterns
- *Generate or simulate; create new texts data or simulate conversations*

These tasks assume that you already have chosen data

Identifying your task

What do you want to do with your text?

- Are you open to learn from your data/you don't have a set theory or RQ to test or answer → discovery
- Do you want to measure how much of property X is present → measurement
- Want to see how respondents write text answer depending on blurb → inference
- *Do you want generate new text data? → generate*

Discovery

Discovery is related to inductive research, and working with text is often an iterative process because of high dimensionality!

- Identify new concepts or structures
- Still guided by theory
- Still need to test and validate patterns that you find

Connected with unsupervised methods

Measurement

Measure to what extent a concept is present in texts to describe society or actors; depend on your research aim

- Identify theoretically relevant concepts to measure
- Train a model to measure that concept (or use output from unsupervised approach)
- Validation is key!

Can be done with both unsupervised, semi-supervised, and supervised methods

Inference/Prediction

Use measures from text to make inference or predict the future

- Have already extracted a measurement from text; use it!
  - Use as dependent variable in regression
  - Use as explanatory variable in regression
- In causal analysis; text can be viewed as the treatment, a confounder, or the outcome

Generation

Generate new texts using LLMs to analyze or simulate conversations between actors in ABM

- Come up with useful prompt with specified task
- Feed LLM your prompt to get output
- Validation is key; currently it is kind off a Wild West

3 challenges with using text-as-data for social science

While computational text analysis holds more promise than ever for social inquiry; some challenges remain

1. Reliability and validity
2. Adopting methods for sociological inquiry
3. Model evaluation and validation

Reliability and validity

Not only methods that need to be reliable and valid but also data

Survey data has standard approaches to evaluate and adjust issues with data quality, no such standards with textual data; e.g.

- Issues when transforming printed material to digital material (e.g. OCR errors)
- Issues with representatives in sample

Issues may increase the variance or bias of estimates

No one-size-fits-all solution; depend (partly) on your research aim

Adopting methods for sociological inquiry

Methods from other fields (NLP & ML) may not adhere to social science ideas

- Where is theory in discovery tasks?
- How to interpret relationships between input and output in neural network models?
- How do we interpret results from pre-trained models (we might not even now what training data they use)?
- Can we replicate the results from unsupervised tasks?

Model evaluation and validation

Many “quatitative measues” to evaluate models focus on convergence or full model performance; while you might only be interested in part of model

Go-to method is to compare model output with human annotators; but how do we validate discoveries?

Many tend to rely on “face validity” (does it make sense?); subjective and many researcher’s degree of freedom

Summary

Text is the most powerful when considered a social sensor of society or actors

Computational text analysis should be used to improve reading; but it does not replace it

Next couple of lectures you will learn about methods but as social scientists; methods should be chosen based on the task; not their novelty!

Let's get started!

www.liu.se

LINKÖPING UNIVERSITY