Topics Methodology: Text as Data
Department of Political Science
POLI:7002
Th 2:00 – 4:50
103 Schaeffer Hall

Instructor	Prof. Colin Case
Email	colin-case@uiowa.edu
Office	359 Schaeffer Hall
Office Hours	Wednesday 11:00-2:00

COURSE DESCRIPTION

In many contexts, the actions of political actors occur in a textual medium (e.g., party manifestos, social media posts, treaties, ordinances, and congressional speeches). However, this data does not come in tidy datasets ready for analysis. This course introduces PhD students to the theory and methods of analyzing textual data within political science and the social sciences. We will focus on converting unstructured text into useful information that can be deployed for scientific inquiry. Topics will include cutting-edge "text as data" tools that allow students to analyze substantive concepts of interest in political texts as part of a research project. We will aim to cover both the conceptual and the practical, including substantial amounts of class time dedicated to working with **R**, an open-source statistical software.

Students should have taken POLI:7003 or have an equivalent stats background. Coding experience in **R** is recommended but not required. Students with little to no experience in **R** should spend extra time early in the semester focusing on DataCamp assignments.

LEARNING OBJECTIVES

1. Understanding the role of text analysis in the social sciences and political science
2. Selecting a corpus, extracting text, and representing text numerically
3. Measuring quantities of interest from text, either using supervised or unsupervised methods
4. Developing coding proficiency in **R**, both related to text analysis and other data science tasks

CLASS STRUCTURE

As an upper-level methods class surveying an advanced methodological topic, we will focus on the breadth, not depth. There are limitless (and growing) tools for text analysis, but not all tools are appropriate for all research questions. As such, you are expected to build a deeper mathematical and foundational understanding of the methods relevant to their research projects and areas of interest. The reading list at the end of the schedule is meant to facilitate this. Each week, there is a mix of conceptual readings (readings that cover principles related to the week's topic; most of this will be found in the textbook for the course), technical readings (readings that introduce the method and the mathematical foundation for the method; covered in both the textbook and additional readings), and applied readings (readings that use the method in an applied way to answer a political science question). On the course calendar, readings are listed roughly in this order.

This class is also designed to ensure students have the tools to carry out original research using text analysis. Classes will be structured to reflect this. In a typical class, there will be about an hour and a half lecture on the topic for the class, a thirty-minute or so coding demonstration on the topic where students will code in **R** alongside the instructor, and the remaining time will be dedicated to students working individually and in groups on an in-class lab. Students must submit their lab before Wednesday at 5 PM the following week on Canvas.

REQUIRED TEXTS

For this course, there is a required textbook. In addition, other readings for the class will be posted on ICON. Students should read the assigned reading in the order presented on the syllabus.

- Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. Text as data: A new framework for machine learning and the social sciences. Princeton University Press, 2022. *(listed as GRS on the calendar)*
- Additional readings posted on ICON.

SOFTWARE

We will be using **R** for all applied work (available for download at `http://cran.rstudio.com/`). I recommend using the RStudio IDE to interact with **R**. Please download and ensure both **R** and RStudio work before the first day of class. While other coding languages can facilitate text analysis (specifically Python), **R** offers broader utility for statistics and data science. There are certain methodological tools only available in Python. If one of those tools is relevant to your research, we can discuss this further to ensure you can conduct the necessary analysis.

COURSE REQUIREMENTS AND GRADING

Grades will be calculated as follows:

A+ 99.00-100
A 93.00-98.99
A- 90.00-92.99
B+ 87.00-89.99
B 84.00-86.99
B- 80.00-83.99
C+ 77.00-79.99
C 74.00-76.99
C- 70.00-73.99
D+ 65.00-69.99
D 64.99-60.00
F 59.99-0

The following components will make up your final grade:

Data Camp Exercises (10%)
To help better develop your R coding skills, you must complete a series of DataCamp assignments in the first part of the semester. Before the semester, I will add you to a DataCamp classroom with a series of modules for you to complete. DataCamp assignments must be finished before class the week they are due. Please note the first DataCamp assignment is due on the first day of classes. All DataCamp modules will be graded as either complete or incomplete.

Pre-Lecture Questions (10%)
To ensure that you get the most out of each class session, you must either a) post at least one question prompted by the study of the reading materials assigned for each week or b) answer at least one of the current questions posted by your classmates. Questions and/or answers should be posted on the corresponding ICON Discussion Topic (labeled by week) no later than 5:00 pm on Tuesday. These will help guide the discussion during our in-person sessions. You are strongly encouraged to answer each others' questions!

Weekly Lab Exercises (30%)
Towards the end of each class, you will work independently or in groups on a practical lab that enables you to apply what you have learned about the week's topic. While there will be time in each class to work on these assignments, you should expect to complete some of the work for each lab outside of class. Labs will be graded as either exceeding expectations (code is complete, commented in detail, and no errors in the code), meeting expectations (code is mostly complete, commented sparingly, and minimal errors in the code), below expectations (code is mostly complete, minimal comments, and numerous errors in the code), or not complete (large sections of the lab are not complete or not submission is made). Graduate student attendance is expected in class, and you will not receive credit for weekly lab exercises

if you do not attend that week's class. Students will only be graded on their 10 highest scores out of 13 total weeks.

Project Poster (50%)
At the end of the semester, you will be required to submit an academic poster of a project that uses one of the methods covered during the semester. You should pursue a project related to your research interests with the potential for academic publication. The following components will make up your grade:

1. Research question and hypothesis (10% of overall grade) **Due February 7th on ICON**: You will submit a one-page memo succinctly outlining what your research question is and what hypothesis(es) you intended to test. This memo should be motivated by substantive questions you are interested in. I strongly encourage you to speak with your advisor about what project you plan to work on early in the semester.

2. Data, measurement, and proposed method (10% of overall grade) **Due March 14th on ICON**: You will submit a 1-2 page memo detailing what data you intend to use for the project and how you plan to obtain it, what quantity of interest you intend to measure in the data, and what method you intend to use. Collecting text data within a semester will be more difficult for some of you than others based on your field of study. If you anticipate collecting your data will be difficult, please meet with me to talk about this more before submitting this assignment.

3. Project Poster and Presentation (30% of overall grade) **Due TBD**: During finals week, you will present a poster that overviews your project. In addition to presenting the main problem to solve or puzzle to study, the poster should include a short discussion of the model's implementation and some preliminary analysis/results. You will be graded on the content of your poster and your short presentation (3-minute elevator pitch). The final presentation will be open to other faculty and grad students at the University, so you are encouraged to invite colleagues to attend.

COMMUNICATION

I am very happy to meet with students outside of class time. Whether it be to discuss concerns about the course, questions about the material, or to engage further with the topic, please feel free to come to office hours. I will be holding office hours in 359 Schaeffer Hall. If you cannot meet during my office hours, which are listed at the top of this syllabus, please email me to set up an alternative time. Office hours are an important resource that should be utilized to improve understanding of the material or ask more personalized questions.

Outside of office hours, e-mail is the easiest way to contact me. I will typically respond to emails within 48 hours. Please send a follow-up if I do not respond to your email within this time frame. Communicating complicated statistical or coding-related questions over email is highly ineffective. Unless the question you are asking is easily expressed in a short paragraph or less, you should plan on attending office hours to ask course-related questions or ask questions during class. I will frequently send emails about the course material, upcoming

assignments or activities, and general reminders. You should check your UI email regularly to stay on top of these updates.

ACADEMIC HONESTY AND MISCONDUCT

Academic dishonesty — including cheating, plagiarism, or any instance of taking credit for work that is not your own — will not be tolerated in this course. All students in CLAS courses are expected to abide by the hrefhttps://clas.uiowa.edu/academics/handbook/standards/academic-honestycollege's standards of academic honesty. Undergraduate academic misconduct must be reported by instructors to CLAS according to these procedures.

You probably have had a chance to work with generative A.I. models such as chat-GPT. These tools are incredibly powerful and can be a huge resource to us as researchers. In fact, we will be using them in later portions of the course to label text data! But just as being exposed to tools like a calculator can prevent you from developing foundational mathematical skills and mental frameworks, so too using pre-trained large language models to come up with answers to assignments can keep you from learning the fundamentals of quantitative research and coding. These models are good, and are getting better, but they can also fail in spectacular (and insidious) ways, and your ability to correct errors generated by the tool will depend on having a solid grasp on said fundamentals. Also, your ability to make the most of these tools is a function of having a strong coding foundation. The more you understand when it comes to coding fundamentals, the better the answers you will get from large language models. I cannot prevent you from using the tool (and I would encourage you to use it responsibly!), and as graduate students, I trust you understand the value of not cheating yourself out of learning fundamental skills in empirical research.

MENTAL HEALTH RESOURCES & STUDENT SUPPORT

Students are encouraged to be mindful of their mental health and seek help as a preventive measure or if feeling overwhelmed and/or struggling to meet course expectations. Students are encouraged to talk to their instructor for assistance with course-related concerns. For additional mental health support, please see the guidance and resources at `mentalhealth.uiowa.edu`, including the 24-7 UI Support and Crisis Line.

Additionally, the Office of the Dean of Students can help students navigate personal crisis situations. They can provide one-on-one support, help with identifying options, and access to basic needs resources (such as food, rent, childcare, etc.). Student Care and Assistance: 132 IMU, dos-assistance@uiowa.edu, or 319-335-1162 and more info: `dos.uiowa.edu/assistance`.

UNIVERSITY POLICIES (LINKS)

- **Accommodations for Students with Disabilities**
- **Free Speech & Expression**

- **Absences for Religious Holidays**
- **Classroom Expectations**
- **Non-discrimination**
- **Sexual Harassment/Misconduct & Supportive Measures**

COURSE SCHEDULE

January 23	Text Analysis in the Social Sciences & R Fundamentals • GRS Ch 1-2 • Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." • Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. "Computer-assisted text analysis for comparative politics." • **Data Camp: Introduction to R and Data Manipulation in R**
January 30	Harvesting Text & Text Representation • GRS Ch 3-4 • Bradley, Alex, and Richard JE James. "Web scraping using R." • **Data Camp: Intermediate R**
February 6	Text Representation & Pre-Processing II • GRS Ch 5-6 • Denny, Matthew J., and Arthur Spirling. "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it." • **Data Camp: Introduction to NLP Ch 1&2 and Introduction to Data Visualization** • **Research question and hypothesis due February 7th on ICON**
February 13	Discovery Principles, Clustering, and Similarity • GRS 7, 10-12 • Carlson, Taylor N. "Through the grapevine: Informational consequences of interpersonal political communication." • **Data Camp: String Manipulation with stringr**
February 20	Principles of Measurement, Counting Methods & Supervised Classification I • GRS 15-17 • Jung, Jae-Hee. "The mobilizing effect of parties' moral rhetoric." • Park, Ju Yeon and Jacob M. Montgomery. "Towards a framework for creating trustworthy measures with supervised machine learning."

February 27	Catch-up and Project Workshop • N/A
March 6	Supervised Classification II • GRS 18-20 • Schub, Robert. "Informing the leader: Bureaucracies and international crises." • Fowler, Erika Franklin, Michael M. Franz, Gregory J. Martin, Zachary Peskowitz, and Travis N. Ridout. "Political advertising online and offline." • Gohdes, Anita R. "Repression technology: Internet accessibility and state violence."
March 13	Supervised Classification III (Active Learning) • Miller, Blake, Fridolin Linder, and Walter R. Mebane Jr. "Active learning approaches for labeling text: review and assessment of the performance of active learning approaches." • Bosley, Mitchell, Saki Kuzushima, Ted Enamorado, and Yuki Shiraito. "Improving Probabilistic Models In Text Classification Via Active Learning." • Kim, Taegyoon. "Violent political rhetoric on Twitter." • **Data, measurement, and proposed method due March 14th on ICON**
March 20	No class, spring break • N/A
March 27	Topic Models I • GRS 13 • "Park, Baekkwan, Amanda Murdie, and David R. Davis. "The (co) evolution of human rights advocacy: Understanding human rights issue emergence over time." • Mueller, Hannes, and Christopher Rauh. "Reading between the lines: Prediction of political violence using newspaper text." • Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. "Structural topic models for open-ended survey responses."
April 3	No Class (MPSA) • Project work day

April 10

Topic Models II

- Ying, Luwei, Jacob M. Montgomery, and Brandon M. Stewart. "Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures."
- Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki. "Keyword-assisted topic models."
- Jankin, Slava, Alexander Baturo, and Niheer Dasandi. "Words to unite nations: The complete United Nations General Debate Corpus, 1946–present."

April 17

Word Embeddings

- GRS 8
- Rodriguez, Pedro L., and Arthur Spirling. "Word embeddings: What works, what doesn't, and how to tell the difference for applied research."
- Kozlowski, Austin C., Matt Taddy, and James A. Evans. "The geometry of culture: Analyzing the meanings of class through word embeddings."
- Esberg, Jane, and Alexandra A. Siegel. "How exile shapes online opposition: Evidence from Venezuela."

April 24

Word Embeddings Extensions

- Rodriguez, Pedro L., Arthur Spirling, and Brandon M. Stewart. "Embedding regression: Models for context-specific description and inference."
- Rodriguez, Pedro, Arthur Spirling, Brandon M. Stewart, and Elisa M. Wirsching. "Multilanguage word embeddings for social scientists: estimation, inference and validation resources for 157 languages."
- Case, Colin R., and Rachel Porter. "Measuring Policy-Level Positioning in US Congressional Elections."

May 1

Transformer Models

- Devlin, Jacob. "Bert: Pre-training of deep bidirectional transformers for language understanding."
- Lin, Gechun. "Using Cross-Encoders to Measure the Similarity of Short Texts in Political Science."
- Widmann, Tobias, and Maximilian Wich. "Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text."
- Wahman, Michael, Nikolaos Frantzeskakis, and Tevfik Murat Yildirim. "From thin to thick representation: How a female president shapes female parliamentary behavior."

May 8	Large Language Models • Ornstein, Joseph T., Elise N. Blasingame, and Jake S. Truscott. "How to train your stochastic parrot: Large language models for political texts." • Barrie, Christopher, Alexis Palmer, and Arthur Spirling. "Replication for Language Models: Problems, Principles, and Best Practice for Political Science." • Valez, Yamil. "Do Personal Issue Priorities Trump Group Policies? Exploring the Impact of Deeply-Held Issues among Latinos using Personalized Conjoint Experiments."
May TBD	Poster Presentations • Location, date, and time to be announced at a later date