Text As Data
PS 2751
Fall 2024
Thursday: 5pm-7:30pm
Room: 4430 Wesley W Posvar Hall

Professor Contact Information

Rebecca Cordell
REC164@pitt.edu
Thursday 3pm-4pm

Course Description

Automated text analysis has become widely used in the social sciences following recent innovations in machine learning and the increased digitalization of political texts. This PhD-level course introduces students to the theoretical underpinnings of text analysis and the different techniques for systematically extracting and analyzing text with applications to political science topics. The focus of the course is on practical applications that allow students to apply cutting edge statistical and computational methods to their own research.

We start the course with the conceptual foundations of quantitative text analysis. We then proceed to consider how we can extract, pre-process, and describe social text data. After that, we explore supervised and unsupervised machine learning methods that can be used to measure and analyze textual content. Students will become familiar with dictionary methods, supervised classification models, sentiment analysis, clustering, structural topic models, word embeddings, and replication and validation.

Active participation in class discussions and computer labs is central to the course. Each week, students will complete the assigned readings, provide insights on the topics, and complete in-class programming activities. Students will develop a general understanding of the text as data literature and will focus in-depth on one method from the course by developing an independent research data paper.

Learning Objectives

Upon completion of this course, students should:

- Learn techniques for measuring and analyzing social science concepts using text as data.
- Develop programming skills in R and a familiarity with text analysis packages.

- Think critically about research design choices and validating measurement models.
- Develop an independent research data paper that applies one of the methods from the course to social text data.

Course Prerequisites

Students should have some familiarity with research design, statistical analysis, and R programming language.

Required Textbooks and Materials

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Copies of the textbook and journal articles are available to students for free via the University of Pittsburgh Library System.

Course Requirements

Assignments

1. **Active Participation** (15% of your grade): Regular attendance and participation in discussions and programming exercises is essential to your success in this course. Research shows that students learn more and develop greater critical thinking skills when they actively participate in their learning. Students are expected to 1) complete the assigned readings before class, 2) provide insights on the topics during discussions in class and online via the discussion board, and 3) work together on interactive programming activities. Students are expected to arrive in class before the start of the class period.

2. **Problem Sets** (25% of your grade): Students will be evaluated on biweekly take-home programming exercises in R. There are a total of six problem sets. For each problem set, students must 1) submit original R code that carries out the required steps, 2) include comments in the code that explain each step, and 3) print your answers to the specified questions in the code. Students are allowed to drop their lowest grade for one problem set.

3. **Research Data Paper** (50% of your grade): Write a 8,000 word original research data paper that applies one of the methods covered during this course to social text data. A good data paper identifies a specific question and puzzle, critically engages with the methods literature, outlines the data

and methods used in the paper, presents and interprets the article's findings, and poses future research questions and investigation. Your final research paper should be conference ready and include the following aspects below.

- Deadline (25% of your grade): December 15.

4. **Research Presentation** (10% of your grade): Give a 15-minute conference style presentation that summarizes your Research Data Paper. You will be graded on the content and quality of your presentation as well as your ability to offer constructive and respectful feedback to other students on how to improve their research projects. The research presentations will take place during the last week of class.

   - Deadline (10% of your grade): December 12.

Late assignments: Written assignments are due on the date assigned, in the form specified. Students who submit their assignments late will have points deducted from their assignment (10% within 1 hour, 25% within 12 hours, 50% within 24 hours, 75% within 48 hours, 100% more than 48 hours). I reserve the right to make exceptions to this policy as circumstances warrant, usually only with prior approval or under instances of extreme emergency or serious illness.

Student Hours: Students are required to meet with me during our student hours twice in the semester. Student hours benefit students academically and professionally by providing you with the opportunity to ask questions about the course content, gain feedback on your research and coursework ideas, and build a professional relationship with the Professor. If you would like me to provide specific feedback on your work during our student hours, you should send via email the relevant materials 48 hours in advance. I am committed to answering your questions and concerns and look forward to getting to know you.

Course Calendar

Week 1 (August 29): What is Text As Data and Why Should We Study it?

Krippendorff, Klaus. 2019. "Conceptual Foundation." *Content Analysis: An Introduction to its Methodology*. SAGE Publications. Chapter 2.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. "Social Science Research and Text Analysis." In *Text as Data: A New Framework for Machine Learning and the Social Sciences.* Princeton University Press. Chapter 2.

Wilkerson, John and Andreu Casas. 2017. “Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges.” *Annual Review of Political Science* 20(1): 529-544.

Week 2 (September 5): No Class

Deadline: Problem Set 1 due on September 08.

Week 3 (September 12): Concepts and Measurement

Krippendorff, Klaus. 2019. “Uses and Inferences.” *Content Analysis: An Introduction to its Methodology*. SAGE Publications. Chapter 3.

Rubin, Donald B. 2008. “For Objective Causal Inference, Design Trumps Analysis.” *Annals of Applied Statistics* 2(3): 808-840.

Shmueli, Galit. 2010. “To Explain or to Predict?” *Statistical Science* 25(3): 289-310.

Week 4 (September 19): Extracting Text

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. “Principles of Selection and Representation” and “Selecting Documents.” In *Text as Data: A New Framework for Machine Learning and the Social Sciences.* Princeton University Press. Chapter 3 & 4.

Mitchell, Ryan. 2018. “Your First Web Scraper.” *Web Scraping with Python: Collecting More Data from the Modern Web*. O’Reilly. Chapter 1.

Munzert, Simon, Christian Rubba, Peter Meißner and Domonic Nyhuis. 2014. “Scraping the Web.” *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. Wiley. Chapter 9.

Deadline: Problem Set 2 due on September 22.

Week 5 (September 26): Pre-processing Text

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. “Bag of Words”, “The Multinomial Language Model”, and “The Vector Space Model and Similarity Metrics.” In *Text as Data: A New Framework for Machine Learning and the Social Sciences.* Princeton University Press. Chapter 5, 6 & 7.

Denny, Matthew J. and Arthur Spirling. 2018. “Text Preprocessing for Unsupervised Learning: Why it Matters, when it Misleads, and what to do About it.” *Political Analysis* 26(2): 168–89.

Herrera, Yoshiko M. and Devesh Kapur. 2007. “Improving Data Quality: Actors, Incentives, and Capabilities.” *Political Analysis* 15(4): 365–86.

Week 6 (October 10): How Can We Describe And Analyze Textual Content?

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. “Representations from Language Sequences”, “Principles of Discovery”, and “Discriminating Words.” In *Text as Data: A New Framework for Machine Learning and the Social Sciences.* Princeton University Press. Chapter 9, 10 & 11.

Ban, Pamela, Alexander Fouirnaies, Andrew B. Hall and James M. Snyder. 2018. “How Newspapers Reveal Political Power.” *Political Science Research and Methods* 7(4): 661–78.

Cordell, Rebecca, K. Chad Clay, Christopher J. Fariss, Reed M. Wood and Thorin M. Wright. 2020. “Changing Standards or Political Whim? Evaluating Changes in the Content of US State Department Human Rights Reports Following Presidential Transitions.” *Journal of Human Rights* 19(1): 3-18.

Deadline: Problem Set 3 due on October 13.

Week 7 (October 17): Dictionary Methods

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. “Principles of Measurement” and “Word Counting.” *Text as Data: A New Framework for Machine Learning and the Social Sciences.* Princeton University Press. Chapter 15 & 16.

Cordell, Rebecca, Kristian Skrede Gleditsch, Florian G. Kern and Laura M. Saavedra-Lux. 2020. “Measuring Institutional Variation across American Indian Constitutions using Automated Content Analysis.” *Journal of Peace Research* 57(6): 777-788.

Shellman, Stephen. 2017. “Coding Disaggregated Intrastate Conflict: Machine Processing the Behavior of Substate Actors Over Time and Space.” *Political Analysis* 16(4): 464-477.

Week 8 (October 24): Supervised Classification Models

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. “An Overview of Supervised Classification”, “Coding a Training Set”, and “Classifying Documents with Supervised Learning.” *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press. Chapter 17, 18 & 19.

Cordell, Rebecca, K. Chad Clay, Christopher J. Fariss, Reed M. Wood and Thorin M. Wright. 2022. “Disaggregating Repression: Identifying Physical Integrity Rights Allegations in Human Rights Reports.” *International Studies Quarterly* 66(2).

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. “How Censorship in China Allows Government Criticism but Silences Collective Expression.” *American Political Science Review* 107(2): 326-343.

Deadline: Problem Set 4 due on October 27.

Week 9 (October 31): Sentiment Analysis

Young, Lori, and Stuart Soroka. 2012. “Affective News: The Automated Coding of Sentiment in Political Texts.” *Political Communication* 29(2): 205-231.

Rudkowsky, Elena, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich and Michael Sedlmair. 2018. “More Than Bags of Words: Sentiment Analysis with Word Embeddings.” *Communication Methods and Measures* 12(2-3): 140-157.

Park, Baekkwan, Kevin Greene, and Mike Colaresi. 2020. "Human Rights are (Increasingly) Plural: Learning the Changing Taxonomy of Human Rights from Large-scale Text Reveals Information Effects." American Political Science Review 114(3): 888-910.

Week 10 (November 7): No Class

Deadline: Problem Set 5 due on November 10.

Week 11 (November 14): Clustering and Structural Topic Models

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. “Clustering” and “Topic Models.” *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press. Chapter 12 and 13.

Terman, Rochelle. 2017. “Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage.” *International Studies Quarterly* 61(2): 489-502.

Huff, Connor and Dominika Kruszewska. 2016. “Banners, Barricades, and Bombs: The Tactical Choices of Social Movements and Public Opinion.” *Comparative Political Studies* 49(13): 1774-1808.

Week 12 (November 21): Word Embeddings

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. “Distributed Representation of Words” and “Low Dimensional Word Embeddings.” *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press. Chapter 8 & 14.

Rodriguez, Pedro and Arthur Spirling. 2021. “Word Embeddings: What Works, What Doesn’t, and How to Tell the Difference for Applied Research.” *Journal of Politics*, 84(1): 101-115.

Hu, Yibo, Mohammad Saleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D’Orazio. 2022. “ConfliBERT: A Pre-trained Language Model for Political Conflict and Violence.” In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States. Association for Computational Linguistics: 5469–5482.

Deadline: Problem Set 6 due on November 24.

Week 13 (November 28): No Class

Week 14 (December 5): Why is Replication and Validation So Important?

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. “Checking Performance” and “Repurposing Discovery Methods” *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press. Chapter 20 & 21.

Arlot, Sylvain and Alain Celisse. 2010. “A Survey of Cross-validation Procedures for Model Selection.” *Statistics Surveys* 3: 40-79.

Dalson, Figueiredo Filho, Rodrigo Lins, Amanda Domingos, Nicole Janz and Lucas Silva. 2019. "Seven Reasons Why: A User’s Guide to Transparency and Reproducibility." *Brazil Political Science Review* 13(2): 1-37.

Week 15 (December 12): Research Presentations

Deadline: Research Data Paper due on December 15.

Course & Instructor Policies

Academic Integrity: Students in this course will be expected to comply with the University of Pittsburgh’s Policy on Academic Integrity. Any student suspected of violating this obligation for any reason during the semester will be required to participate in the procedural process, initiated at the instructor level, as outlined in the University Guidelines on Academic Integrity. This may include, but is not limited to, the confiscation of the examination of any individual suspected of violating

University Policy. Furthermore, no student may bring any unauthorized materials to an exam, including dictionaries and programmable calculators.

Disability Services: If you have a disability for which you are or may be requesting an accommodation, you are encouraged to contact both your instructor and Disability Resources and Services (DRS), 140 William Pitt Union, (412) 648-7890, drsrecep@pitt.edu, (412) 228-5347 for P3 ASL users, as early as possible in the term. DRS will verify your disability and determine reasonable accommodations for this course.

Content Warning and Class Climate Statement: Our course readings and classroom discussions will often focus on mature, difficult, and potentially challenging topics. As with any course in political science, course topics are often political and personal. Readings and discussions might trigger strong feelings—anger, discomfort, anxiety, confusion, excitement, humor, and even boredom. Some of us will have emotional responses to the readings; some of us will have emotional responses to our peers' understanding of the readings; all of us should feel responsible for creating a space that is both intellectually rigorous and respectful. Above all, be respectful (even when you strongly disagree) and be mindful of the ways that our identities position us in the classroom.

I expect everyone to come to class prepared to discuss the readings in a mature and respectful way. If you are struggling with the course materials, here are some tips: read the syllabus so that you are prepared in advance. You can approach your instructor ahead of time if you'd like more information about a topic or reading. If you think a particular reading or topic might be especially challenging or unsettling, you can arrive to class early and take a seat by the door so that you can easily exit the classroom as needed. If you need to leave or miss class, you are still responsible for the work you miss. If you are struggling to keep up with the work because of the course content, you should speak with me and/or seek help from the counseling center.

Equity, Diversity, and Inclusion: The University of Pittsburgh does not tolerate any form of discrimination, harassment, or retaliation based on disability, race, color, religion, national origin, ancestry, genetic information, marital status, familial status, sex, age, sexual orientation, veteran status or gender identity or other factors as stated in the University's Title IX policy. The University is committed to taking prompt action to end a hostile environment that interferes with the University's mission. For more information about policies, procedures, and practices, visit the Civil Rights & Title IX Compliance web page.

I ask that everyone in the class strive to help ensure that other members of this class can learn in a supportive and respectful environment. If there are instances of the aforementioned issues, please contact the Title IX Coordinator, by calling 412-648-7860, or e-mailing titleixcoordinator@pitt.edu. Reports can also be filed online. You may also choose to report this to a faculty/staff member; they are

required to communicate this to the University's Office of Diversity and Inclusion. If you wish to maintain complete confidentiality, you may also contact the University Counseling Center (412-648-7930).

Email Communication: Each student is issued a University e-mail address (username@pitt.edu) upon admittance. This e-mail address may be used by the University for official communication with students. Students are expected to read e-mail sent to this account on a regular basis. Failure to read and react to University communications in a timely manner does not absolve the student from knowing and complying with the content of the communications. The University provides an e-mail forwarding service that allows students to read their e-mail via other service providers (e.g., Hotmail, AOL, Yahoo). Students that choose to forward their e-mail from their pitt.edu address to another address do so at their own risk. If e-mail is lost as a result of forwarding, it does not absolve the student from responding to official communications sent to their University e-mail address.

Gender Inclusive Language Statement: Language is gender-inclusive and non-sexist when we use words that affirm and respect how people describe, express, and experience their gender. Just as sexist language excludes women's experiences, non-gender-inclusive language excludes the experiences of individuals whose identities may not fit the gender binary, and/or who may not identify with the sex they were assigned at birth. Identities including trans, intersex, and genderqueer reflect personal descriptions, expressions, and experiences. Gender-inclusive/non-sexist language acknowledges people of any gender (for example, first year student versus freshman, chair versus chairman, humankind versus mankind, etc.). It also affirms non-binary gender identifications, and recognizes the difference between biological sex and gender expression. Students, faculty, and staff may share their preferred pronouns and names, and these gender identities and gender expressions should be honored.

Religious Observances: The observance of religious holidays (activities observed by a religious group of which a student is a member) and cultural practices are an important reflection of diversity. As your instructor, I am committed to providing equivalent educational opportunities to students of all belief systems. At the beginning of the semester, you should review the course requirements to identify foreseeable conflicts with assignments, exams, or other required attendance. If at all possible, please contact me (your course coordinator/s) within the first two weeks of the first class meeting to allow time for us to discuss and make fair and reasonable adjustments to the schedule and/or tasks.

Sexual Misconduct, Required Reporting, and Title IX: If you are experiencing sexual assault, sexual harassment, domestic violence, and stalking, please report it to me and I will connect you to University resources to support you.

University faculty and staff members are required to report all instances of sexual misconduct, including harassment and sexual violence to the Office of Civil Rights

and Title IX. When a report is made, individuals can expect to be contacted by the Title IX Office with information about support resources and options related to safety, accommodations, process, and policy. I encourage you to use the services and resources that may be most helpful to you.

As your professor, I am required to report any incidents of sexual misconduct that are directly reported to me. You can also report directly to Office of Civil Rights and Title IX: 412-648-7860 (M-F; 8:30am-5:00pm) or via the Pitt Concern Connection at: Make A Report

An important exception to the reporting requirement exists for academic work. Disclosures about sexual misconduct that are shared as a relevant part of an academic project, classroom discussion, or course assignment, are not required to be disclosed to the University's Title IX office.

If you wish to make a confidential report, Pitt encourages you to reach out to these resources:

- The University Counseling Center: 412-648-7930 (8:30 A.M. TO 5 P.M. M-F) and 412-648-7856 (AFTER BUSINESS HOURS)
- Pittsburgh Action Against Rape (community resource): 1-866-363-7273 (24/7)

If you have an immediate safety concern, please contact the University of Pittsburgh Police, 412-624-2121.

Any form of sexual harassment or violence will not be excused or tolerated at the University of Pittsburgh.

Statement on Classroom Recording: To ensure the free and open discussion of ideas, students may not record classroom lectures, discussion and/or activities without the advance written permission of the instructor, and any such recording properly approved in advance can be used solely for the student's own private use.

Statement on Scholarly Discourse: In this course we will be discussing very complex issues of which all of us have strong feelings and, in most cases, unfounded attitudes. It is essential that we approach this endeavor with our minds open to evidence that may conflict with our presuppositions. Moreover, it is vital that we treat each other's opinions and comments with courtesy even when they diverge and conflict with our own. We must avoid personal attacks and the use of ad hominem arguments to invalidate each other's positions. Instead, we must develop a culture of civil argumentation, wherein all positions have the right to be defended and argued against in intellectually reasoned ways. It is this standard that everyone must accept in order to stay in this class; a standard that applies to

all inquiry in the university, but whose observance is especially important in a course whose subject matter is so emotionally charged.

Your Well-being Matters: College/Graduate school can be an exciting and challenging time for students. Taking time to maintain your well-being and seek appropriate support can help you achieve your goals and lead a fulfilling life. It can be helpful to remember that we all benefit from assistance and guidance at times, and there are many resources available to support your well-being while you are at Pitt. You are encouraged to visit Thrive@Pitt to learn more about well-being and the many campus resources available to help you thrive.

If you or anyone you know experiences overwhelming academic stress, persistent difficult feelings and/or challenging life events, you are strongly encouraged to seek support. In addition to reaching out to friends and loved ones, consider connecting with a faculty member you trust for assistance connecting to helpful resources.

The University Counseling Center is also here for you. You can call 412-648-7930 at any time to connect with a clinician. If you or someone you know is feeling suicidal, please call the University Counseling Center at any time at 412-648-7930. You can also contact Resolve Crisis Network at 888-796-8226. If the situation is life threatening, call Pitt Police at 412-624-2121 or dial 911.

The descriptions and timelines contained in this syllabus are subject to change at the discretion of the Professor.