Text Analysis for Public Policy

Fall 2025

Instructor:	Daniel Weitzel	**Time:**	F 2PM–5PM
Email:	daniel.weitzel@colostate.edu	**Place:**	Clark C347

Objectives:

I am excited to welcome you to **Text Analysis and Natural Language Processing (NLP) for Public Policy**, a course designed to equip you with the tools and techniques needed to analyze large-scale textual data in the social sciences. Over the semester, we will explore a series of modules that introduce key concepts in computational text analysis, from fundamental text preprocessing to advanced machine learning applications. No prior programming experience is required – early sessions will provide an introduction to `R`, ensuring that everyone is comfortable working with text as data.

The course follows a structured progression, beginning with **text preprocessing techniques**, including tokenization, stemming, lemmatization, and part-of-speech tagging. From there, we will explore **dictionary-based methods, topic modeling, and sentiment analysis**, applying these techniques to real-world policy documents, legislative texts, and social media data. As we advance, we will cover **word embeddings, supervised text classification, and even transformer-based models**, focusing on their applications in policy research, public opinion analysis, and media studies. Throughout the course, we will emphasize the ethical implications of automated text analysis, addressing issues such as algorithmic bias and data privacy.

This course will challenge you, but it is also an opportunity to expand your skill set as a policy researcher. You will gain hands-on experience in text analysis, work on practical policy-related case studies, and develop your ability to extract meaningful insights from unstructured text. The course will culminate in a poster session, where you will apply the skills you have learned to a policy-relevant research question, analyze a large text dataset, and present your findings to your peers. This final project will allow you to showcase your ability to use text-as-data methods to inform public policy decisions.

I encourage you to ask questions, engage in discussions, and collaborate with your peers. Together, we will explore how computational text analysis can enhance policy research and decision-making in the digital age.

Office Hours:

Monday 12-3pm in Clark B348. I am also available after class, or by appointment. During office hours I am available for any and all questions students might have. Please make use of this opportunity. We can discuss your questions about the course material, the class, or your research project.

Semester Overview

This course provides a comprehensive introduction to text analysis techniques and their application in the realm of public policy. The semester is structured around a series of practical and theoretical sessions that equip students with the skills to collect, preprocess, analyze, and interpret large volumes of text data to inform policy decisions. Key topics covered include classical text analysis methods, machine learning approaches, and cutting-edge technologies such as BERT and Large Language Models.

Weeks 1-4: Foundations of Text Analysis

The semester begins with an introduction to the role of text analysis in public policy and key challenges in working with text data. Students will explore methods for collecting and preprocessing text data from various sources, learning the essentials of data collection, cleaning, and tokenization. The focus will be on gaining proficiency in R programming for text analysis and understanding the ethical considerations when working with text data. By the end of Week 4, students will have developed the skills to start working with descriptive analysis methods like word frequency analysis, and dictionary-based approaches such as sentiment and topic tagging.

Weeks 5-8: Advanced Text Analysis Techniques

In Weeks 5 through 8, the course dives deeper into sentiment analysis, opinion mining, and topic modeling. Students will learn to apply Latent Dirichlet Allocation (LDA) for uncovering topics in large text corpora. The analysis will expand to comparing texts across different groups, time periods, and sources to identify policy trends and differences. Additionally, scaling methods will be introduced to handle large datasets, ensuring that students can work efficiently with text data at scale.

Weeks 9-11: Machine Learning Approaches

Weeks 9 through 11 focus on introducing machine learning techniques for text analysis. Students will learn both supervised and unsupervised machine learning approaches, such as classification and clustering methods, to extract insights from text data. They will also be introduced to advanced models like BERT and RoBERTa, which leverage deep learning for more nuanced understanding of text. These methods are applied to real-world public policy problems, preparing students for advanced text analysis.

Weeks 12-14: Cutting-Edge Techniques and Student Projects

In the final weeks, students will explore Large Language Models (LLMs), including small pretrained models such as Microsoft's Phi4 and Deepseek, alongside more sophisticated models like BERT and RoBERTa. These tools will be explored in the context of their application to public policy analysis. The course concludes with a poster session in Week 14, where students will present their final projects, demonstrating their ability to apply text-as-data techniques to analyze real-world public policy issues.

Readings

This course will utilize one primary textbook along with a variety of academic articles that apply text-as-data (TADA) approaches to social science questions. Each week, we will also review additional articles that illustrate the application of TADA methods in the social sciences. These readings will be made available on the corresponding module pages on Canvas.

GRS Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as data: A new framework for machine learning and the social sciences. Princeton University Press.

In addition, the following books can allow you to dive deeper into specific text analysis topics. If you want to really master a specific text-as-data approach these books can serve as a starting point. They include natural language processing specific topics as well as broader introduction to machine learning.

- Krippendorff, K., 2018. Content analysis: An introduction to its methodology. Sage publications.
- Manning, C. D., Raghavan, P. & Schütze, H. (2008). Introduction to Information Retrieval, Cambridge University Press.
- Jurafsky, D., & Martin, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.
- Jurafsky, D. & Martin, J. (2008).Speech and Language Processing, 2nd Edition. Prentice Hall.
- Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning. New York: Springer.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. New York: Springer.

The programming language used in this class is `R`. If you wish to deepen your understanding of data transformation and visualization in R, the following freely available resources are highly recommended:

- Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. 2023. R for data science. O'Reilly Media, Inc.
- Healy, K. 2018. Data visualization: A practical introduction. Princeton University Press.
- Silge, J. 2017. Text mining with R: A tidy approach. O'Reilly Media, Inc.

Suggested Reading:

At the end of the syllabus, you will find an extensive list of academic articles that employ text analysis methods to address social science questions. While this list is not required reading, it is provided as a resource for inspiration and guidance as you develop your own research projects. When you begin formulating your research ideas, feel free to explore this list to find examples of methodologies and applications that align with your interests. You don't have to reinvent the wheel and your poster can start out as a replication.

Evaluation

Class Participation (10%)

Active participation in class discussions is critical to the success of this course. To contribute effectively, you must complete all assigned readings before class and actively engage in discussions. High-quality participation means staying on topic, being well-prepared, and offering insightful contributions to the class. Engaging with the course content, asking questions, and seeking clarification on difficult concepts also contribute positively to participation. If you are unsure about a particular concept, chances are your peers may have similar questions, so raising them benefits everyone.

Homework Assignments (30%)

There will be **thirteen short homework assignments** throughout the semester, contributing to 30% of your final grade. These assignments are designed to reinforce class content and ensure you can apply the methods discussed in class. They will primarily focus on **real-world applications** of the techniques learned, allowing you to gradually build up the skills necessary for your final project. The homeworks should guide you toward your research poster by helping you practice implementing methods such as data collection, preprocessing, and analysis techniques. These assignments will vary in difficulty, with each progressively building toward your poster project. The homeworks will cover areas like **text preprocessing**, **descriptive text analysis**, **machine learning approaches**, and specific techniques used in your final project.

Research Poster (60%)

The final project is a **research poster** based on your application of **text-as-data analysis (TADA)** methods to a **public policy question**. The poster should present your **research question**, **hypotheses**, **methods**, and **results** in a visually accessible format. This project is broken down into three parts:

1. **Replication Files (25%)**
   Submit the **replication files** for the analysis you present in your research poster. This includes all scripts, cleaned data, and code used to reproduce the results. The goal is to demonstrate that your findings are replicable and based on sound methodology.

2. **Initial and Final Research Poster (45%)**
   The initial poster should summarize the scope and direction of your project, while the final poster should present completed research with results and conclusions. **Weekly check-ins** (see below) will help guide you as you refine your project. Your poster should clearly outline your policy question, methods, and findings, presented in an accessible and engaging manner.

   - **Week 9**: **Poster Proposal** submission and meeting with the instructor for feedback and guidance.
   - **Weeks 10-11**: Reports on **poster progress**, detailing your analysis, challenges, and next steps.
   - **Week 12**: **Poster check-in** with draft posters for feedback.
   - **Week 13**: **Poster check-in** with draft posters, finalizing content and design before the presentation.

3. **Research Poster Presentation and Q&A (30%)**
   The final component of your research poster grade is the **presentation**. During the poster session, you will present your project to your peers and engage in a **Q&A** session. This is an opportunity to demonstrate your understanding of the analysis, respond to feedback, and explain how your work contributes to the broader field of public policy.

Class Policies

- **Attendance** Regular attendance is essential for your academic success. I expect that you attend and actively participate in all class sessions. Absence or lateness does not excuse students from required course work. This is a graduate seminar, which means that I expect that you come to class prepared with notes about the reading and ready to actively engage in class discussion.

- **Communication** The most reliable way to get in touch with me is via email. You should expect a response within 48 business hours. Please write emails that are efficient and to the point.

- **Academic Honesty and Integrity** This course will adhere to the CSU Academic Integrity Policy as found on the Student' Responsibilities page of the CSU General Catalog and in the Student Conduct Code. At a minimum, violations will result in a grading penalty in this course and a report to the Office of Student Resolution Center. Lack of knowledge of the academic honesty policy is not a reasonable explanation for a violation.

- **Accommodations** Your experience in this class is important to me. If you require any accommodation, let me know ahead of time what would be helpful so that we can plan together for you to succeed. You do not need to share private information with me, but you must provide verifiable documentation to the Office of Student Case Management or Student Disability Center. For religious accommodations, please complete the Religious Accommodation Request Form. Please provide verifiable documentation to them (not to me!) ahead of time and ensure that they forward me this information *at least one week* prior to the assignment for which accommodations are required. I cannot make adjustments after the fact.

- **Late Assignments** I will not accept any late assignments. Exceptions are granted only if the Office of Student Case Management is able to provide documentation of a health emergency or other life emergency. If you experience an emergency, please contact Student Case Management, which will then contact me.

- **Grievances** If you are unhappy with your grade on an assignment, please wait 24 hours after the assignment is returned before contacting me. This provides the opportunity to let the initial emotions subside and think more clearly about the issue at hand. After 24 hours, you can contact me with a written explanation of why you feel your grade should be different. "I worked hard" is not a good explanation; I can only grade the quality of the work that you give to me.

- **Intellectual Growth** The goal of this class is the personal and intellectual growth of all students. Every student is expected to participate in the generation of an respectful and professional environment that facilitates this growth.

  – Woolley, Kaitlin, and Ayelet Fishbach. 2022. "Motivating personal growth by seeking discomfort." *Psychological Science*, 33.4: 510-523.

- **Mental Health** Feeling like a big failure and worried that everybody will find out? Guess what, you are not alone! Imposter syndrom is very common in graduate school. There are ways we as a class can help each others. Be respectful and mindful in how you interact with your colleagues. Always contribute to an open and engaging class environment. I strongly encourage students to ask questions. If you don't understand something, you are usually not alone. There are mental health resources available online and on campus. You can also read more about the issue here:

  – Almasri, N., Read, B. and Vandeweerdt, C., 2022. "Mental Health and the PhD: Insights and Implications for Political Science." *PS: Political Science & Politics*, 55(2), pp.347-353.

Week 1: Introduction to Text Analysis for Public Policy

This session introduces the role of text analysis in public policy research and decision-making. We will explore key challenges in working with text data, including issues of scale, structure, and ethical considerations. Students will also gain foundational skills in `R` programming for text analysis, learning how to set up reproducible research workflows and implement ETL (Extract, Transform, Load) pipelines for text data collection and preprocessing. By the end of the session, students will understand the theoretical foundations of text analysis and be equipped with practical tools to begin working with text as data in public policy research.

Learning Goals

- Understand the significance of text analysis in public policy research.
- Identify key challenges in handling text data, including noise, structure, and ethical concerns.
- Gain familiarity with `R` packages for text analysis (e.g., `tm`, `quanteda`, `tidytext`).
- Learn best practices for structuring reproducible text analysis projects.
- Develop basic ETL pipelines for collecting, cleaning, and processing text data.

Topics Covered

- Introduction to text analysis and NLP techniques for public policy.
- Ethical considerations in text analysis (privacy, bias, and transparency).
- Setting up an `R` programming environment for text analysis.
- Organizing project structures using RMarkdown and version control.
- Implementing ETL pipelines: data collection (APIs, web scraping), preprocessing (cleaning, tokenization, transformation), and workflow integration.
- Overview of project expectations and poster presentations.

Readings

1. **Theoretical:** One of the following:
    - GRS, Chapter 1.
    - Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21(3),* 267-297.
2. **Theoretical:** O'Connor, B., Bamman, D., & Smith, N. A. (2011). Computational text analysis for social science: Model assumptions and complexity. *Second workshop on computational social science and the wisdom of crowds (NIPS 2011).*
3. **Application:** Jin, Z., & Mihalcea, R. (2023). Natural language processing for policymaking. In *Handbook of Computational Social Science for Policy* (pp. 141-162). Springer.

Week 2-4: Text Data Collection and Preprocessing

This session focuses on acquiring and preparing text data for analysis. We will explore various sources of text data, including social media, news articles, and legislative texts, and discuss methods for collecting text using web scraping and APIs. Students will also learn essential preprocessing techniques, such as tokenization, stopword removal, stemming, and lemmatization, to clean and normalize text for analysis. By the end of this session, students will be able to collect, preprocess, and structure text data using `R`.

Learning Goals

- Understand different sources of text data relevant to public policy research.
- Learn methods for web scraping and API data collection (e.g., Wikipedia, news websites).
- Apply key text preprocessing techniques, including tokenization, stopwords removal, stemming, and lemmatization.
- Implement text normalization techniques such as case folding and punctuation removal.
- Gain hands-on experience collecting and processing text data in `R`.

Topics Covered

- Introduction to data sources: social media, news articles, legislative texts, etc.
- Methods for web scraping and API usage.
- Text preprocessing: cleaning, tokenization, stopwords, stemming, and lemmatization.
- Hands-on: Collecting and processing text data in `R`.

Readings

1. **Theoretical:** One of the following:
   - GRS Chapter 4-5
   - Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis, 26(2)*, 168-189.
2. **Computational:** Silge, J., Robinson, D., & Robinson, D. (2017). The tidy text format. In *Text Mining with R: A Tidy Approach*. O'Reilly.
3. **Application:** Understanding the Potential of Text Mining for Equity Analysis by the Urban Institute.
4. **Application:** One of the following:
   - Cross, J. P., & Hermansson, H. (2017). Legislative amendments and informal politics in the European Union: A text reuse approach. *European Union Politics*, 18(4), 581-602.
   - O'Connor, B., Stewart, B. M., & Smith, N. A. (2013, August). Learning to extract international relations from political context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 1094-1104).
5. **Recommended:** Wikipedia on Text encoding, `http://en.wikipedia.org/wiki/Text_encoding`

Week 5: Descriptive Text Analysis and Dictionary Methods

This session introduces students to fundamental descriptive text analysis techniques and dictionary-based methods. We will explore how to quantify text characteristics using word frequency analysis, n-grams, and word clouds. Additionally, we will cover dictionary-based approaches such as topic tagging and word associations, which allow researchers to classify and interpret textual data systematically. By the end of this session, students will be able to apply these techniques in R to analyze public policy texts.

Learning Goals

- Understand word frequency analysis and its applications in text data.
- Learn how to implement dictionary-based methods such as topic tagging and word associations.
- Explore the strengths and limitations of dictionary-based approaches.
- Use R packages (`tm`, `quanteda`, `tidytext`) for descriptive text analysis.

Topics Covered

- Word frequency analysis: token counts, n-grams, and word clouds.
- Introduction to dictionary-based methods: topic tagging and word associations.
- Using R packages like `tm`, `quanteda`, and `tidytext` for descriptive analysis.

Readings

1. **Theoretical:** GRS Chapter 16
2. **Theoretical:** Benoit, K., Laver, M., & Mikhaylov, S. (2009). Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science, 53(2)*, 495-513.
3. **Application:** One of the following:
    - Spirling, A. (2016). Democratization and linguistic complexity: The effect of franchise extension on parliamentary discourse, 1832–1915. *The Journal of Politics, 78(1)*, 120-136.
    - Hengel, E. (2022). Publishing while female: Are women held to higher standards? Evidence from peer review. *The Economic Journal, 132(648)*, 2951-2991.

Week 6: Sentiment Analysis and Opinion Mining

This session introduces sentiment analysis techniques to measure and interpret opinions in text data. We will explore different approaches to sentiment analysis, including dictionary-based and machine-learning methods, and discuss their applications in public policy research. By the end of this session, students will be able to apply sentiment analysis techniques to evaluate public opinion and policy-related texts.

Learning Goals

- Understand the foundations of sentiment analysis and its role in public policy research.
- Learn dictionary-based and machine-learning approaches to sentiment classification.
- Implement sentiment analysis techniques using R packages (`tidytext`, `quanteda`, `textdata`).
- Critically assess the limitations and biases in sentiment analysis models.
- Conduct a hands-on sentiment analysis of public policy texts.

Topics Covered

- Introduction to sentiment analysis and opinion mining.
- Dictionary-based sentiment analysis: positive/negative word scoring.
- Machine-learning approaches to sentiment classification.
- Hands-on: Analyzing the sentiment of public policy texts.

Readings

1. **Theoretical:** Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2(1–2)*, 1-135.
2. **Computational:** Silge, J., Robinson, D., & Robinson, D. (2017). Sentiment analysis with tidy data . In *Text Mining with R: A Tidy Approach.* O'Reilly.
3. **Application:** One of the following:
    - Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication, 29(2)*, 205-231.
    - Dodds, P. S., & Danforth, C. M. (2009). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies, 11(4)*, 441-456.
4. **Recommended:** What Yankee Candle reviews can tell us about COVID

Week 7: Topic Modeling and Latent Dirichlet Allocation (LDA)

This session introduces topic modeling techniques for uncovering hidden themes in large text corpora. We will cover probabilistic topic models, with a focus on Latent Dirichlet Allocation (LDA) and its extensions. Special attention will be given to seeded LDA, which incorporates prior information to improve topic coherence and interpretability. By the end of this session, students will be able to apply topic modeling techniques to policy-related texts and evaluate model outputs.

Learning Goals

- Understand the principles of topic modeling and its applications in public policy research.
- Learn how to implement LDA and seeded LDA in R.
- Evaluate and interpret topic model results.
- Explore methods for selecting the number of topics and assessing model quality.

Topics Covered

- Introduction to topic modeling: Unsupervised learning for text classification.
- Latent Dirichlet Allocation (LDA): Assumptions, priors, and hyperparameters.
- Seeded LDA: Incorporating domain knowledge to guide topic modeling.
- Model evaluation: Perplexity, coherence scores, and human validation.
- Hands-on: Applying topic modeling to public policy texts in R.

Readings

1. **Theoretical:** GRS Chapter 13
2. **Theoretical:** Mohr, J. W., Bogdanov, P. (2013). *Introduction—Topic models: What they are and why they matter. Poetics, 41(6)*, 545-569.
3. **Application:** One of the following:
    - Eshima, S., Imai, K., & Sasaki, T. (2024). Keyword-assisted topic models. *American Journal of Political Science*, 68(2), 730-750.
    - Anastasopoulos, L. J., Moldogaziev, T., & Scott, T. (2017). Computational text analysis for public management research. Available at SSRN 3269520.

Week 8: Comparing Texts Across Groups, Time, and Sources

This session covers methods for systematically comparing text corpora across different groups, time periods, and sources. We will explore quantitative measures of textual similarity, word usage differences, and changes in discourse over time. By the end of this session, students will be able to compare text datasets effectively and interpret meaningful differences.

Learning Goals

- Understand different approaches to comparing texts across categories.
- Learn methods for quantifying textual differences.
- Explore temporal analysis of text and changes in discourse over time.
- Apply comparison techniques to public policy texts.

Topics Covered

- Comparing word frequency distributions: cosine similarity, Jaccard similarity.
- Measuring linguistic and topical shifts over time.
- Identifying key differences in text corpora using keyword analysis (e.g., TF-IDF).
- Using structural topic models (STM) for dynamic topic comparisons.
- Hands-on: Analyzing differences in political discourse across time and sources.

Readings

1. **Theoretical:** Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis, 16(4)*, 372-403.
2. **Application:** One of the following:
    - Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis, 18(1)*, 1-35.
    - Eshbaugh-Soha, M. (2010). The tone of local presidential news coverage. *Political Communication, 27(2)*, 121-140.

Week 9: Scaling Methods for Text Data

This session introduces scaling methods that allow researchers to place texts on a continuous dimension, such as ideology, sentiment, or policy position. We will cover supervised and unsupervised scaling approaches and discuss their applications in public policy research. By the end of this session, students will be able to apply scaling techniques to extract meaningful latent traits from text data.

Learning Goals

- Understand the theoretical foundations of text scaling.
- Learn and compare supervised and unsupervised scaling techniques.
- Implement text scaling in R and interpret results.
- Critically assess the validity and assumptions of different scaling methods.

Topics Covered

- Introduction to text scaling: What does it mean to "scale" a text?
- Unsupervised scaling methods:
  - Wordfish: A Poisson-based model for estimating positions from text.
  - Wordshoal: Scaling texts that vary over time or across groups.
- Supervised scaling methods:
  - Wordscores: Scaling texts based on reference documents.
  - Machine learning approaches for text classification and scaling.
- Applications of text scaling in public policy research.
- Hands-on: Estimating ideological positions from legislative speeches.

Readings

1. **Theoretical:** Laver, M., Benoit, K., & Garry, J. (2003). *Estimating policy positions from political texts using words as data. American Political Science Review, 97(2)*, 311-331.
2. **Theoretical:** Lowe, W., Benoit, K., Mikhaylov, S., & Laver, M. (2011). *Scaling policy preferences from coded political texts. Legislative Studies Quarterly, 36(1)*, 123-155.
3. **Application:** Slapin, J. B., & Proksch, S.-O. (2008). *A scaling model for estimating time-series party positions from texts. American Journal of Political Science, 52(3)*, 705-722.

Week 10: Machine Learning – Supervised Approaches to Text Analysis

This session covers supervised machine learning methods for text analysis, where labeled training data is used to build predictive models. We will explore different classification and regression techniques commonly applied in public policy research. By the end of this session, students will be able to train, evaluate, and interpret supervised machine learning models for text classification.

Learning Goals

- Understand the fundamentals of supervised machine learning for text analysis.
- Learn key classification algorithms, including logistic regression, decision trees, and support vector machines.
- Implement text classification using bag-of-words and word embeddings.
- Evaluate model performance using accuracy, precision, recall, and F1-score.
- Discuss ethical considerations and potential biases in supervised learning.

Topics Covered

- Introduction to supervised learning: Text classification and regression.
- Feature engineering for text data: Bag-of-words, TF-IDF, and word embeddings.
- Classification models:
  - Logistic regression and Naïve Bayes classifiers.
  - Decision trees and random forests.
  - Support vector machines (SVMs) and deep learning approaches.
- Hands-on: Training and evaluating text classifiers in R.

Readings

1. **Theoretical:** GSM Chapter 19
2. **Theoretical:** Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (Chapters on text classification).
3. **Application:** Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Machine Learning for Social Science: An Introduction*. Cambridge University Press.
4. **Application:** Taddy, M. (2013). *Multinomial inverse regression for text analysis. Journal of the American Statistical Association, 108(503)*, 755-770.

Week 11: Machine Learning – Unsupervised Approaches to Text Analysis

This session covers unsupervised machine learning techniques that help uncover patterns, structures, and relationships in text data without labeled training examples. We will explore clustering, dimensionality reduction, and representation learning methods. By the end of this session, students will be able to use these methods to analyze and interpret text data in new ways.

Learning Goals

- Understand the principles of unsupervised machine learning for text.
- Learn clustering techniques such as k-means and hierarchical clustering.
- Implement word embeddings to uncover semantic relationships in text.
- Apply dimensionality reduction techniques to visualize text data.
- Evaluate the interpretability and reliability of unsupervised methods.

Topics Covered

- Introduction to unsupervised learning: When and why to use it.
- Clustering techniques for text data:
  - k-means clustering and hierarchical clustering.
  - DBSCAN and spectral clustering.
- Word embeddings and representation learning:
  - Word2Vec, GloVe, and FastText embeddings.
  - Analyzing word similarity and semantic relationships.
- Dimensionality reduction:
  - Principal Component Analysis (PCA).
  - t-SNE and UMAP for text visualization.
- Hands-on: Clustering documents and exploring semantic spaces.

Readings

1. **Theoretical:** GSM Chapter 8
2. **Theoretical:** Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
3. **Application:** Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review, 84(5)*, 905-949.

Week 12: BERT and RoBERTa for Text Analysis

This session delves into contextual embeddings through transformer models, with a focus on BERT and RoBERTa. These models allow for more accurate semantic analysis and are key tools for tasks like text classification, named entity recognition, and question answering.

Learning Goals

- Understand the architecture and capabilities of BERT and RoBERTa models.
- Learn how to fine-tune these models for specific text classification tasks.
- Explore the practical applications of BERT and RoBERTa in public policy and other domains.
- Understand the key differences between BERT and RoBERTa.

Topics Covered

- Introduction to BERT and RoBERTa:
  - How transformers work and their advantage over traditional models.
  - Differences between BERT and RoBERTa.
- Fine-tuning BERT and RoBERTa:
  - Using Hugging Face's 'transformers' library to fine-tune BERT and RoBERTa.
  - Pretraining and transfer learning with BERT and RoBERTa for domain-specific tasks.
- Applications:
  - Text classification, named entity recognition, and sentiment analysis.
  - Policy text analysis using BERT and RoBERTa.

Readings

1. **Theoretical:** Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2024). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis*, 32(1), 84-100.
2. **Application:** Widmann, T., & Wich, M. (2023). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. Political Analysis, 31(4), 626-641.
3. **Application:** Timoneda, J. C., & Vera, S. V. (2025). BERT, RoBERTa or DeBERTa? Comparing Performance Across Transformers Models in Political Science Text. *The Journal of Politics*.

Week 13: Large Language Models (LLMs) and Small Pretrained Models

This session covers the latest advancements in large language models (LLMs), focusing on both open-source models and smaller, efficient alternatives that you can actually run locally. We will explore how to use Ollama for local inference and evaluate compact, open-source LLMs like Microsoft's Phi-4 and DeepSeek.

Learning Goals

- Understand the architecture and capabilities of modern LLMs.
- Learn about smaller, open-source LLMs and their trade-offs.
- Deploy and fine-tune models locally using Ollama and other frameworks.
- Explore LLM applications in text classification, summarization, and policy analysis.

Topics Covered

- Introduction to Large Language Models:
  - How transformers scale (GPT, LLaMA, Mixtral, etc.).
  - Strengths and weaknesses of proprietary vs. open-source models.
- Compact Open-Source Models:
  - Microsoft's Phi-4, DeepSeek, and other small-scale LLMs.
  - Trade-offs in size, performance, and fine-tuning.
- Running and Deploying LLMs:
  - Local inference using Ollama.
  - Efficient prompt engineering and fine-tuning.
  - Hardware considerations for running models locally.
- Hands-on:
  - Running LLMs on personal machines.
  - Using small LLMs for text classification and summarization.
  - Comparing outputs from different models on policy-related texts.

Readings

1. **Theoretical:** Ornstein, J. T., Blasingame, E. N., & Truscott, J. S. (2023). How to train your stochastic parrot: Large language models for political texts. *Political Science Research and Methods*, 1-18.
2. **Theoretical:** Wu, P. Y., Nagler, J., Tucker, J. A., & Messing, S. (2023). Large language models can be used to estimate the latent positions of politicians. arXiv preprint arXiv:2303.12057.
3. **Application:** Microsoft's Phi-4 Model Overview.
4. **Application:** Ollama Documentation on Local LLM Inference.

Week 14: Poster Session - Student Projects on Text as Data in Public Policy

In this session, students will present their projects applying text-as-data techniques to real-world public policy issues. Through a poster presentation format, students will share the methods, findings, and policy implications of their work. The focus will be on how text analysis can inform public policy decisions, identify trends, and support evidence-based policymaking.

Suggested Readings

The articles below are a (incomplete) list of applications of text as data approaches to social science questions with public policy implications. This list is meant as a resource for you to find a starting point for your own research project. Go through the list of articles and select articles either for their methodology or their topic. Your work on your research poster can start with a replication of an existing project!

- Alvarez, R.M. and Morrier, J., 2024. Evaluating the Quality of Answers in Political Q&A Sessions with Large Language Models. *arXiv preprint arXiv:2404.08816*.
- Arold, B.W., Ash, E., MacLeod, W.B. and Naidu, S., 2024. Do words matter? the value of collective bargaining agreements. *Center for Law & Economics Working Paper Series*, 6.
- Bali, V.A. and Higgins, D., 2023. More than Meets the Eye? Using Text Analytic Techniques to Unpack School Mission Statements. *SAGE Open*, 13(4).
- Barari, S. and Simko, T., 2023. LocalView, a database of public meetings for the study of local politics and policy-making in the United States. *Scientific Data*, 10(1).
- Bauer, M., Huber, D., Offner, E., Renkel, M. and Wilms, O., 2024. Corporate green pledges (No. 214). *IMFS Working Paper Series*.
- Blackington, C. and Cayton, F., 2024. To Dog-Whistle or to Bark? Elite Communication Strategies When Invoking Conspiracy Theories. *Government and Opposition*.
- Bruinsma, B. and Johansson, M., 2024. Finding the structure of parliamentary motions in the Swedish Riksdag 1971–2015. *Quality & Quantity*, 58(4).
- Buylova, A., Nasiritousi, N., Duit, A., Reischl, G. and Lejon, P., 2024. Paper tiger or useful governance tool? Understanding long-term climate strategies as a climate governance instrument. *Environmental Science & Policy*, 159.
- Campiglio, E., Deyris, J., Romelli, D. and Scalisi, G., 2023. Warning words in a warming world: Central bank communication and climate change. *Global Research Alliance for Sustainable Finance and Investment*.
- Cantone, G.G., Tomaselli, V. and Mazzeo, V., 2024. Review bombing: ideology-driven polarisation in online ratings: The case study of The Last of Us (part II). *Quality & Quantity*.
- Caramani, D., Gurova, S. and Widmann, T., 2024. The Evolution of Global Cleavages: A Historical Analysis of Territorial and Functional World Alignments Based on Automated Text Analysis, 1843–2020. *Comparative Political Studies*.
- Chalmers, A., Klingler-Vidra, R. and Malou van den Broek, O., 2024. From diffusion to diffuse-ability: A text-as-data approach to explaining the global diffusion of Corporate Sustainability Policy. *International Studies Quarterly*, 68(1).
- Chen, Y., Long, J., Jun, J., Kim, S.H., Zain, A. and Piacentine, C., 2023. Anti-intellectualism amid the COVID-19 pandemic: The discursive elements and sources of anti-Fauci tweets. *Public Understanding of Science*, 32(5).
- Chiu, S.H., Han, T., Post, A.E., Ratan, I. and Soga, K., 2025. Studying tech adoption with "text-as-data": Opportunities, pitfalls, and complementarities in the case of transportation. *Environment and Planning B: Urban Analytics and City Science*.

- Cuesta-Delgado, D., Barberá-Tomás, D. and Marques, P., 2024. A text-mining analysis of Latin America Universities' mission statements from a 'Third Mission'perspective. *Studies in Higher Education*.
- Dent, C.M., 2024. The UK's new free trade agreements in the Asia-Pacific: how closely is it adopting US trade regulation?. *The Pacific Review*, 37(3).
- Duxbury, S.W., 2024. Collaborating on the Carceral State: Political Elite Polarization and the Expansion of Federal Crime Legislation Networks, 1979 to 2005. *American Sociological Review*.
- Franchino, F., Migliorati, M., Pagano, G. and Vignoli, V., 2024. Concepts and measures of bureaucratic constraints in European Union laws from hand-coding to machine-learning. *Regulation & Governance*, 18(3).
- Graham, R., Schoonvelde, M. and Swinkels, M., 2024. Unpacking the European Commission's fiscal policy response to crisis: mapping and explaining economic ideas in the European Semester 2011–2022. *Journal of European Public Policy*, 31(11).
- Grajzl, P. and Murrell, P., 2024. Caselaw and England's economic performance during the Industrial Revolution: Data and evidence. *Journal of Comparative Economics*, 52(1).
- Gordillo, D.D., Timoneda, J. and Vera, S.V., 2024. Machines Do See Color: A Guideline to Classify Different Forms of Racist Discourse in Large Corpora. *arXiv preprint arXiv:2401.09333*.
- Goutsmedt, A. and Fontan, C., 2023. The ECB and the inflation monsters: strategic framing and the responsibility imperative (1998-2023). *Journal of European Public Policy*.
- Goutsmedt, A. and Fontan, C., 2024. The ECB and the inflation monsters: strategic framing and the responsibility imperative (1998–2023). *Journal of European Public Policy*, 31(4).
- Goutsmedt, A., Sergi, F., Claveau, F. and Fontan, C., 2023. The Different Paths of Central Bank Scientization: The Case of the Bank of England.
- Hunter, T. and Walter, S., 2025. International organizations in national parliamentary debates. *The Review of International Organizations*.
- Hurtado Bodell, M., Magnusson, M. and Keuschnigg, M., 2024. Seeded topic models in digital archives: Analyzing interpretations of immigration in Swedish newspapers, 1945–2019. *Sociological Methods & Research*.
- Ishima, H., 2024. Talking Like Opposition Parties? Electoral Proximity and Language Styles Employed by Coalition Partners in a Mixed Member Majoritarian System. *Legislative Studies Quarterly*, 49(3).
- Jamison, A., Hoelscher, K., Miklian, J., Henisz, W. and Ganson, B., 2024. Is media sentiment associated with future conflict events?. *Authorea Preprints*.
- Juhász, R., Lane, N., Oehlsen, E. and Pérez, V.C., 2022. The who, what, when, and how of industrial policy: A text-based approach. *What, When, and How of Industrial Policy: A Text-Based Approach (August 15, 2022)*.
- Knapp, A., 2024. Protection Trinity: Assessing the Three-tier Framework in United Nations Resolutions. *International Peacekeeping*, 31(4).
- Litofcenko, J., Vogler, A., Meyer, M. and Mehrwald, M., 2023. From controversy to common ground: The discourse of sustainability in the media. *Journal of Language and Politics*, 22(5).

- Li, B., Song, Y., Shi, Y. and Chen, H.T., 2024. Unpacking the complexity of online incivility: an analysis of characteristics and impact of uncivil behavior during the Hong Kong protests. *Internet research*.
- Macanovic, A. and Przepiorka, W., 2023. The moral embeddedness of cryptomarkets: text mining feedback on economic exchanges on the dark web. *Socio-Economic Review*.
- Martin, M.V., Kirsch, D.A. and Prieto-Nañez, F., 2023. The promise of machine-learning-driven text analysis techniques for historical research: topic modeling and word embedding. *Management & Organizational History*, 18(1).
- Matchett, L., 2024. Putting on the Blitz: Urgency and Department of Defense Communications in Times of Budget Shortfall. *Armed Forces & Society*, 50(3).
- Mervaala, E., 2025. Climate Change Versus Economic Growth: Quantifying, Identifying and Comparing Articulations in News Media Using Dynamic Topic Modeling. *Environmental Communication*.
- Mesquita, R., 2024. What do I need to say to get your signature? Adding draft resolution text to the UN General Assembly Sponsorship Dataset. *Research & Politics*, 11(4).
- Morandell, T., Wicki, M. and Kaufmann, D., 2025. The planning of urban–rural linkages: An automated content analysis of spatial plans adopted by European intermediate cities. *Landscape and Urban Planning*, 255.
- Moreno-Cabanillas, A., Castillero-Ostio, E. and Serna-Ortega, Á., 2024. Digital disinformation strategies of European climate change obstructionist think tanks. *Frontiers in Communication*, 9.
- Moreira Ramalho, T., Massart, T. and Crespy, A., 2024. Resilient austerity? National economic discourses before the pandemic in the European Union. *Politics & Policy*, 52(5).
- Oldac, Y.I. and Olivos, F., 2025. The development of higher education research topics between 2000 and 2021: Seven patterns from generalist journals. *Review of Education*, 13(1).
- Ojo, A., Rizun, N., Walsh, G., Mashinchi, M.I., Venosa, M. and Rao, M.N., 2024. Prioritising national healthcare service issues from free text feedback–A computational text analysis & predictive modelling approach. *Decision Support Systems*, 181.
- Righetti, N. and Bertuzzi, N., 2024. Rethinking Veganism in the Digital Age. Innovating Methodology and Typology to Explore a Decade of Facebook Discourses. *Sociological Research Online*.
- Ruan, T. and Lv, Q., 2022. Public perception of electric vehicles on reddit over the past decade. *Communications in Transportation Research*, 2.
- Ruan, T. and Lv, Q., 2023. Public perception of electric vehicles on Reddit and Twitter: A cross-platform analysis. *Transportation research interdisciplinary perspectives*, 21.
- Ruan, T. and Lv, Q., 2024. Exploring equity perception of electric vehicles from a social media perspective. *Transportation Research Interdisciplinary Perspectives*, 25.
- Silva-Muller, L. and Sposito, H., 2024. Which Amazon problem? Problem-constructions and transnationalism in Brazilian presidential discourse since 1985. *Environmental Politics*, 33(3).
- Skov, M. and Svarre, T., 2024. A diachronic cluster analysis of Danish museum websites. *Internet Histories*.

- Squires, A., Clark-Cutaia, M., Henderson, M.D., Arneson, G. and Resnik, P., 2022. "Should I stay or should I go?" Nurses' perspectives about working during the Covid-19 pandemic's first wave in the United States: A summative content analysis combined with topic modeling. *International Journal of Nursing Studies*, 131.
- Steigenberger, N., Garz, M. and Cyron, T., 2024. Signaling theory in entrepreneurial fundraising and crowdfunding research. *Journal of Small Business Management*.
- Struthers, C.L. and Ritzler, C., 2024. Advocacy strategies in state preemption: The case of energy fuel bans. *Policy studies journal*, 52(2).
- Trexler, A., 2024. Aimed emotions in American presidential politics. *Journal of Information Technology & Politics*, 21(4).
- Truscott, J.S., 2024. Analyzing the Rhetoric of Supreme Court Confirmation Hearings. *Journal of Law and Courts*, 12(1).
- Voltmer, J.B., Fisseler, B., Raimann, J. and Stürmer, S., 2024. Using topic modeling to research student diversity in higher education. *Zeitschrift für Psychologie*.
- Vandenbroucke, S., Kantorowicz, J. and Erkens, Y., 2024. Decoding supplier codes of conduct with content and text as data approaches. *Corporate Social Responsibility and Environmental Management*, 31(1).
- Vignoli, V. and Coticchia, F., 2024. The politics of military assistance: Italian parties' positions on the war in Ukraine. *South European Society and Politics*.
- Walter, M., Bagozzi, B.E., Ajibade, I. and Mondal, P., 2023. Social media analysis reveals environmental injustices in Philadelphia urban parks. *Scientific Reports*, 13(1).
- Waldhof, G., 2024. Moral value conflicts in the German debate about genetically engineered foods. *Journal of Consumer Protection and Food Safety*.
- Zhou, A., Liu, W. and Yang, A., 2024. Politicization of science in COVID-19 vaccine communication: Comparing US politicians, medical experts, and government agencies. *Political Communication*, 41(4).
- Zinnatullin, A., 2023. Political discussions in online oppositional communities in the non-democratic context. *Computational Communication Research*, 5(1).