
Making Sense of Text

An Introduction to Text-as-Data Methods for Social Scientists

Instructor

Dr. Melike Ayşe Kocacık Şenol
Sabancı University

Contact

makocacik@sabanciuniv.edu

Course Description

We live in an era of unprecedented textual abundance. Political actors communicate through speeches, manifestos, legislative records, press releases, and social media on a scale that no individual researcher can read systematically. At the same time, the rapid development of computational methods—from classical bag-of-words models to word embeddings and large language models—has fundamentally changed what is possible. Text-as-data approaches are no longer a niche specialization; they are becoming a core competency for empirical social science research.

This course offers a hands-on introduction to this evolving landscape. Participants will work through the complete text-as-data pipeline: acquiring texts, transforming them into structured data, and analyzing them using a range of techniques, while building a critical understanding of what these methods can and cannot do. Special attention is paid to newly emerging approaches, including word embeddings and transformer-based models, alongside established methods such as dictionary analysis and topic modeling. All exercises use the R programming language, with an emphasis on reproducibility and interpretive rigor.

Text-as-data methods are developing rapidly, and the field continues to evolve in ways that no single syllabus can fully capture. The readings listed here represent established, peer-reviewed contributions, but participants are strongly encouraged to track emerging work in outlets such as *Political Analysis*, the *Journal of Politics*, and proceedings from NLP and computational social science venues. One of the core skills this course aims to cultivate is the ability to engage with new methods critically—understanding their assumptions and limitations—rather than adopting them as black boxes.

Course Objectives

By the end of this course, students will be able to:

- situate text-as-data methods within the broader landscape of computational social science, in-

cluding an awareness of how the field is evolving with large language models and transformer-based approaches;

- acquire, clean, and preprocess textual data in R, including tokenization, stopword removal, and construction of document-feature matrices;
- apply supervised, semi-supervised, and unsupervised classification techniques, including dictionary analysis, sentiment analysis, and topic modelling;
- understand the conceptual foundations of word embeddings and contextual representations, and recognize their implications for political text analysis;
- critically evaluate the assumptions, limitations, and potential misapplications of automated text analysis methods;
- communicate findings effectively and transparently, with explicit reflection on methodological choices.

Course Format

The course is held over five days. Each day consists of a morning lecture session and an afternoon practice session. Lectures introduce conceptual and theoretical content; practice sessions provide structured, hands-on engagement with R. Assigned readings should be completed before the corresponding morning session. Students are expected to bring a laptop with R and RStudio installed to each afternoon session.

Schedule

Day	Session	Topic	Content
Day 1	Morning	Introduction to Text-as-Data	What are we doing here? Overview of the course. Key concepts: corpus, document, token, type. Why text matters for political and social science.
	Afternoon	<i>Practice Session 1</i>	Introduction to R and RStudio. Basic data structures. Reading and writing text files.
Day 2	Morning	The Bag of Words Approach	Preprocessing: tokenization, stopword removal, stemming and lemmatization. Document-feature matrices (DFMs). Weighting schemes (TF-IDF).
	Afternoon	<i>Practice Session 2</i>	Building a DFM with <code>quanteda</code> . Visualizing word frequencies. Hands-on with political text corpora.

Day	Session	Topic	Content
Day 3	Morning	Word Embeddings	Limitations of sparse representations. Word2Vec, GloVe, and contextual embeddings. Pre-trained models and their use in political research.
	Afternoon	Practice Session 3	Using word embedding models in R. Measuring semantic similarity. Applied examples.
Day 4	Morning	Supervised & Semi-Supervised Classification	Dictionary analysis: construction, validation, limitations. Sentiment analysis: lexicon-based vs. machine learning approaches. Coding political texts.
	Afternoon	Practice Session 4	Sentiment scoring with existing dictionaries. Custom dictionary construction. Evaluating classifier performance.
Day 5	Morning	Unsupervised Classification Techniques	Topic models: LDA and structural topic models (STM). Word counts and scaling methods. Interpreting and validating results.
	Afternoon	Practice Session 5	Running LDA with <code>topicmodels</code> and <code>stm</code> . Tuning the number of topics. Presenting results from unsupervised models.
Day 6	Morning	From Embeddings to LLMs: The Transformer Revolution	The attention mechanism and transformer architecture. BERT, GPT, and domain-specific models (e.g. ConflBERT, PoliBERT). Zero-shot, few-shot, and fine-tuned classification. Opportunities and risks for social science research.
	Afternoon	Practice Session 6	Querying LLMs via API for text classification tasks. Prompt design for political text. Comparing LLM outputs with classical methods. Critical discussion: validation, reproducibility, and the black-box problem.

Assigned Readings by Day

Day	Readings
Day 1	<p>Grimmer, J. and Stewart, B.M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. <i>Political Analysis</i>, 21(3), 267–297.</p> <p>Benoit, K. (2020). Text as data: An overview. In: Curini, L. and Franzese, R. (eds.), <i>The SAGE Handbook of Research Methods in Political Science and International Relations</i>. SAGE, pp. 461–497.</p>
Day 2	<p>Grimmer, J., Roberts, M.E. and Stewart, B.M. (2022). <i>Text as Data: A New Framework for Machine Learning and the Social Sciences</i>. Princeton University Press. Chapter 4: Bag of Words.</p> <p>Quanteda Initiative (2022). quanteda: Quantitative Analysis of Textual Data. R package documentation. https://quanteda.io</p>
Day 3	<p>Rodriguez, P.L. and Spirling, A. (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. <i>Journal of Politics</i>, 84(1), 101–115.</p> <p>Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space. <i>arXiv preprint arXiv:1301.3781</i>.</p>
Day 4	<p>Young, L. and Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. <i>Political Communication</i>, 29(2), 205–231.</p> <p>Laver, M., Benoit, K. and Garry, J. (2003). Extracting policy positions from political texts using words as data. <i>American Political Science Review</i>, 97(2), 311–331.</p>
Day 5	<p>Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet allocation. <i>Journal of Machine Learning Research</i>, 3, 993–1022.</p> <p>Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B. and Rand, D.G. (2014). Structural topic models for open-ended survey responses. <i>American Journal of Political Science</i>, 58(4), 1064–1082.</p>
Day 6	<p>Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. <i>Proceedings of NAACL-HLT 2019</i>, 4171–4186.</p> <p>Törnberg, P. (2023). How to use large language models for text analysis. <i>arXiv preprint arXiv:2307.13106</i>.</p> <p>Ornstein, J.T., Case, E.A. and Hammond, J.S. (2025). How to train your stochastic parrot: Large language models for political texts. <i>Political Science Research and Methods</i>, 13(2), 264–281.</p>