GESIS Leibniz Institute for the Social Sciences

gesis training

GESIS Fall Seminar in Computational Social Science 2025
"From Embeddings to LLMs: Advanced Text Analysis with Python"

Lecturer: Hauke Licht
Affiliation: University of Innsbruck
Email: hauke.licht@uibk.ac.at

Date: September 22-26, 2025
Time: 9:30-12:30 and 13:30-16:30

About the Lecturer

Hauke Licht is assistant professor of Computational Political Science at the Department of Political Science and the Digital Science Center, University of Innsbruck. He received his PhD from the University of Zurich and previously worked as a post-doctoral researcher at the University of Cologne. In his research, Hauke develops and applies deep learning-based computational text analysis methods to study political communication, electoral competition, and democratic representation with a focus on multilingual analyses and text annotation.

Course Description

Basic "bag-of-words" methods of text analysis treat words or *n*-grams as distinct symbols and texts as unordered collections of such symbols. This inherently limits bag-of-words methods' ability to represent many of the nuances and subtilities of natural language that make studying text so interesting for social scientists. Deep learning methods for text embedding and neural language modeling help overcome the limitations of bag-of-words text analysis approaches and thus are an essential addition to the toolkit of computational social science researchers.

This course introduces social scientists to advanced, deep learning-based text analysis methods. Participants will learn about the conceptual motivation and methodological foundations of text embedding methods and large neural language models (LLMs). Moreover, they will gather plenty of practical experience with applying these methods in social science research using the Python programming language. Next to conveying a solid conceptual understanding as well as hands-on experience with applying these methods, the course puts a strong emphasis on introducing and discussing potential social science use cases as well as ethical considerations.

- We will start by discussing *text embedding methods*, beginning with an overview of static word embedding models like the GloVe and word2vec algorithms and followed by contextualized embeddings and the Transformer architecture. Participants will learn how to use embeddings for document search and clustering using the `scikit-learn` and `sentence-transformers` Python packages.
- We will then cover the methodological foundations of state-of-the-art pre-trained *masked language models* like BERT and introduce model finetuning. Participants will learn to apply pre-trained models in *supervised learning* tasks (single- and multilabel sentence classification, token classification) using the `transformers` and `setfit` packages and *topic modelling* with `BERTopic`.
- Next, we will focus on *generative language models* like GPT and the foundations of large language models (LLMs). Participants will learn techniques to prompt LLMs to analyze and annotate texts based on no or only a few labelled examples (i.e., zero-shot prompting and few-shot in-context learning) and how to implement these techniques using `ollama` and `llama-index` in Python.

This is an advanced-level course. Participants should have prior knowledge of basic text analysis techniques. Specifically, they should have experience with standard bag-of-words pre-processing techniques and text representation approaches, such as word count-based document-feature matrices. Those looking for a more introductory-level course should consider taking "Introduction to Machine Learning for Text Analysis with Python"

(15-19 September). Moreover, participants should have experience with programming in Python. The lecturer cannot introduce or repeat basics in Python programming in the course due to limited time.

Organizational Structure of the Course

The course will be organized as a mixture of lectures and exercise sessions. We will switch between lectures and exercises throughout the morning and afternoon sessions of the course. In the lecture sessions, I will focus on explaining core concepts and methods. In the exercise sessions, participants will apply their newly acquired knowledge. The lecturer will be available to answer questions and provide guidance during the entire course.

Keywords

computational text analysis, word embeddings, large language models (LLMs), Transformers, deep learning, Python

Target Group

You will find the course useful if:

- you have a background in the social sciences or humanities (e.g., communication science, economics, political science, sociology, or related fields)
- you have a solid understanding of basic text analysis methods and
- you want to advance your knowledge, skills, and practical experience
- you want to get up to speed with applying state-of-the-art NLP methods to text analysis problems in social science research

Course and Learning Objectives

By the end of the course, you will:

- know the methodological foundations of text embedding methods, transfer learning, Transformers, large language models (LLMs)
- be able to apply these methods to analyze social scientific text data
- be able to reflect critically on the application of the techniques in social science research, including relevant ethical considerations

Course Prerequisites

- Prior knowledge of basic quantitative text analysis methods
    - bag-of-words text pre-processing ("tokenization") and representation (i.e., how to represent documents with word count vectors)
    - (conceptual) knowledge of dictionary analysis, topic modeling, and supervised text classification methods is strongly recommended
- Basic knowledge of Python
    - creating and manipulating strings, lists and dictionaries
    - creating and interacting with objects, classes and methods
    - reading and manipulating data frames with pandas
    - using loops
    - defining new functions
- Basic knowledge of quantitative research methods
    - understanding of linear and logistic regression analysis
    - a basic understanding of matrix algebra might be helpful but is not required

For those who would like a primer or refresher in Python, we recommend taking the online workshop "Introduction to Python" (25-28 August) and/or the online blended learning course "Introduction to Computational Social Science with Python" (01-05 September).

Software and Hardware Requirements

- Participants must bring their own laptops to this course.
- Participants must have Python (≥ 3.11), miniconda, pip, and Jupyter Notebook installed on their laptops.
- Required Python libraries
  - text processing: `nltk`, `genism`, `transformers`, `setfit`, `sentence-transformers`, `BERTopic`, `llama-index`, `ollama`, `llama-index-llms-ollama`
  - others: `numpy`, `scipy`, `pandas`, `scikit-learn`
- The lecturer will distribute concrete instructions for the Python setup and a comprehensive list of required libraries before the course and assist with any remaining setup problems on the first day of the course.
- It is recommended that participants create a Google Colab account, especially if their laptop has no Nvidia GPU (Windows/Linux) or Apple Silicon chip (MacOS).

Course Contents

- Deep learning-based NLP methods for computational social science text analysis
- Classic (static) word embedding methods
- Contextualized embedding, attention mechanisms, and Transformer models
- Transformer fine-tuning for supervised text classification
- LLM prompting and in-context learning for text annotation

Day-to-day Schedule and Literature

Day 1: Introduction to text embedding methods

We will begin by reviewing the limitations of count-based bag-of-words document representations (insensitivity to words' context, high dimensionality, and sparsity) and the methodological intuition that motivates embedding-based alternatives. We will then cover classic (static) word embedding methods – focusing on word2vec – and learn how to perform basic computations such as finding similar terms, computing words' similarities, and inducing conceptual dimensions.

Next, we will discuss the limitations of static word embedding models, focusing on the issues that arise for words with multiple senses and words whose meaning depends on sentence context. In the context of this discussion, we will introduce the attention mechanism as a technique for computing word context-sensitive embeddings to overcome the multiple word senses problem of "traditional" word embeddings.

During the exercise time, the lecturer will be available for troubleshooting technical issues participants encounter when trying to set up their Python environment.

Recommended literature:

- Rodriguez, P. L., & Spirling, A. (2021). Word Embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*, *84*(1), 101–115. https://doi.org/10.1086/715162
- Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing* (3rd edition). Published online. Chapter 6
- 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. https://doi.org/10.1177/0003122419877135
- Gennaro, G., & Ash, E. (2022). Emotion and Reason in Political Language. *The Economic Journal*, *132*(643), 1037–1059. https://doi.org/10.1093/ej/ueab104
- Hargrave, L., & Blumenau, J. (2022). No Longer Conforming to Stereotypes? Gender, Political Style and Parliamentary Debate in the UK. *British Journal of Political Science, 52*(4), 1584–1601. https://doi.org/10.1017/S0007123421000648
- Rodman, E. (2020). A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. *Political Analysis*, *28*(1), 87–111. https://doi.org/10.1017/pan.2019.23

- Rheault, L., & Cochrane, C. (2020). Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis*, *28*(1), 112–133. https://doi.org/10.1017/pan.2019.26

Day 2: Transformers, BERT, and fine-tuning

Building on the topics covered on the first day, we will introduce the Transformer architecture, focusing on encoder-only masked language models like BERT. We will explore how Transformer models can be used in social science research and then focus on text classification applications. The lecturer will explain and illustrate how to fine-tune a pre-trained Transformer encoder model for classification tasks and how to evaluate classification performance. Participants will then be provided with various labeled text data sets they can use learn to apply supervised fine-tuning techniques for text classification.

Recommended literature:

- Wankmüller, S. (2022). Introduction to Neural Transfer Learning with Transformers for Social Science Text Analysis. *Sociological Methods & Research, 53*(4), 1676-1752. https://journals.sagepub.com/doi/full/10.1177/00491241221134527
- Timoneda, J. C., & Vallejo Vera, S. (2025). BERT, RoBERTa or DeBERTa? Comparing Performance Across Transformers Models in Political Science Text. *The Journal of Politics, 87*(1), 347-364. https://www.journals.uchicago.edu/doi/full/10.1086/730737
- Wang, Y. (2024). On Finetuning Large Language Models. *Political Analysis, 32*(3), 379-383. https://doi.org/10.1017/pan.2023.36
- Do, S., Ollion, É., & Shen, R. (2022). The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy. *Sociological Methods & Research*, *53*(3), 1167-1200. https://doi.org/10.1177/00491241221134526
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. https://arxiv.org/abs/1706.03762
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. https://aclanthology.org/N19-1423/

Day 3: (More) Fun with encoder models

On the third day, we will continue to explore how Transformer models can be used in social science research by learning about methods to fine-tune encoder models to assign texts to more than one category (multilabel classification), extract spans of words from texts (token classification), or pairwise compare texts. Moreover, participants will learn how efficient sentence-transformer fine-tuning with `setfit` enables supervised learning with few labeled examples. We will then shift our focus to the task of topic modeling and introduce the BERTopic framework. Participants will, again, be provided with text data sets they can use to learn to apply this approach in Python.

Recommended literature:

- Erlich A., Dantas S. G., Bagozzi B. E., Berliner D., Palmer-Rubin B. (2022). Multi-Label Prediction for Political Text-as-Data. *Political Analysis, 30*(4), 463-480. https/doi.org/10.1017/pan.2021.15
- Licht, H., & Sczepanski, R. (2024). Who are they talking about? Detecting mentions of social groups in political texts with supervised learning. *British Journal of Political Science,* forthcoming. https://doi.org/10.31219/osf.io/ufb96
- 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. https://doi.org/10.1017/pan.2023.20
- Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient Few-Shot Learning Without Prompts. *arXiv preprint.* https://arxiv.org/abs/2209.11055
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXix preprint*. https://arxiv.org/abs/2203.05794

Day 4: Generative Pre-trained Transformer (GPT) models, LLMs, and in-context learning

On day four, we will catch up with the most recent developments in the computational text analysis social science literature by focusing on using large language models (LLMs) and their use in social science text analysis. We will first learn about the methodological differences between encoder-only models like BERT and generative language models like GPT. We will then discuss the ingredients for LLM development, covering pre-training, instruction tuning, and reinforcement learning with human feedback (RLHF). Next, we will learn about "prompting" LLMs, that is, giving them instructions and how we can use examples, demonstrating the desired response to the model enables "in-context learning."

We will apply these techniques to instruct open-weights LLMs like LLaMa, Phi, Mistral, Gwen, etc. to complete the text classification tasks introduced on days 2 and 3. This will allow for a head-to-head comparison between the supervised fine-tuning and LLM prompting approaches to text classification and give participants concrete examples of best practices in prompting and implementation.

Recommended literature:

- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018) Improving language understanding by generative pre-training.https://gwern.net/doc/www/s3-us-west-2.amazonaws.com/d73fdc5ffa8627bce44dcda2fc012da638ffb158.pdf
- Ziegler, D. M., Stiennon, N. , Wu, J., Brown, T. B. , Radford, A., Amodei, D. , Christiano, P., & Irving, G. (2020). Fine-Tuning Language Models from Human Preferences. https://doi.org/10.48550/arXiv.1909.08593
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). "ChatGPT outperforms crowd workers for text-annotation tasks". In: *Proceedings of the National Academy of Sciences 120*(30), 1-3. https://doi.org/10.1073/pnas.2305016120
- Burnham M. (2024). Stance detection: a practical guide to classifying political beliefs in text. *Political Science Research and Methods*. https://doi.org/10.1017/psrm.2024.35
- Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2024). Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics*, *11*(1). https://doi.org/10.1177/20531680241231468
- Hugging Face (2025): LLM prompting guide. https://huggingface.co/docs/transformers/en/tasks/prompting

Day 5: Advanced LLM prompting

We will follow up on the previous day by adding depth and breadth. First, we will learn how to make LLMs robust text annotation programs by (a) enforcing JSON response formatting and (b) decomposing complex annotation tasks into smaller ones and using "prompt chaining" to compose an LLM workflow. Second, we will add breadth by reviewing the wide range of current and potential uses of LLMs for text analysis and generation in the social sciences. In the afternoon of day 5, the lecturer will meet with participants in 1-on-1 sessions to discuss their research projects, answer questions, etc. on an individual, on-demand basis.

Recommended literature:

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. https://dl.acm.org/doi/abs/10.1145/3442188.3445922
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint* arXiv:2108.07258
- Voelkel, J. G., & Willer, R. (2023). *Artificial Intelligence Can Persuade Humans on Political Issues.* https://osf.io/stakv/
- Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv preprint* https://arxiv.org/abs/2301.04246
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681-694. https://link.springer.com/article/10.1007/s11023-020-09548-1