GESIS Leibniz Institute for the Social Sciences

gesis training

GESIS Fall Seminar in Computational Social Science 2024
"From Embeddings to LLMs: Advanced Text Analysis with Python"

Lecturers:	Hauke Licht	Lisa Maria Lechner
Affiliation:	University of Cologne	University of Innsbruck
Email:	hauke.licht@wiso.uni-koeln.de	lisa.lechner@uibk.ac.at

Date: September 23-27, 2024
Time: 9:30-12:30 and 13:30-16:30

About the Lecturers

Hauke Licht is a post-doctoral researcher at the Cologne Center for Comparative Politics, University of Cologne, and has received his PhD from the University of Zurich. He develops and applies deep learning-based computational text analysis methods to study political communication, electoral competition, and democratic representation. He also has a strong focus on multilingual analyses.

Lisa Maria Lechner is an assistant professor for methods and methodology in political science at the University of Innsbruck. In her research, Lisa studies international treaties, such as trade agreements, bilateral tax treaties, and environmental agreements, as well as national and international jurisdictions by dint of inferential network- and quantitative text analysis. At the moment, she is particularly interested in multilingual word embeddings.

Course Description

Basic "bag-of-words" methods of text analysis that rely on counting words or *n*-grams are limited in their ability to account for the complexity of natural language. This has implications for our ability to apply these approaches to measure social science concepts in textual data. 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 thus introduces social scientists to advanced, deep learning-based text analysis methods such as word embeddings and large neural language models. 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 introducing classical word embedding models like GloVe and word2vec and participants will learn how to use word embeddings in social science research. We will then introduce state-of-the-art Transformer models like BERT and GPT. We will first cover their methodological foundations: the attention mechanism, masked and autoregressive language modeling, and the neural network architectures that characterize BERT and GPT. Participants will then apply these models in exercises covering various supervised learning tasks (single- and multilabel sentence classification, token classification, and pairwise comparison) as well as topic modeling with BERTopic. Finally, we will introduce strategies and techniques to prompt pre-trained generative language models to code texts based on no or only a few labelled examples (i.e., zero-shot prompting and few-shot in-context learning).

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" (16-20 September). Moreover, participants should have experience with programming in Python. The instructors cannot provide an introduction to or recap of 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, we will focus on explaining core concepts and methods. In the exercise sessions, participants will apply their newly acquired knowledge. Both instructors 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 about 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 document 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" (26-29 August) and/or the online blended learning course "Introduction to Computational Social Science with Python" (30 August-05 September).

Software and Hardware Requirements

- Participants should bring their own laptops for use in the course.
- You should have Python (≥ 3.12), miniconda, and Jupyter Notebook installed (see this link).
- Required Python libraries
    - text processing: nltk, scikit-learn, genism, tokenizers, datasets, transformers, openai, sentence-transformers, BERTopic, setfit
    - others: numpy, scipy, pandas
- The instructors 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.
- Parts of the exercises focusing on LLM prompting techniques will require participants to (i) sign up for an account with a commercial provider (OpenAI) and (ii) add credit to their account. However, the instructors will ensure that the costs for using commercial providers' models will remain below U.S. $ 10. Moreover, the instructors will present open-source alternatives participants will be able to use free of charge through Google Colab or locally on their computers and laptops. The relevant information and setup instructions will be shared with registered participants 4 weeks in advance of the course to allow the instructors to adapt to the currently rapid evolution of available open-source models and software solutions.

Day-to-day Schedule and Literature

Day 1: Introduction to classic word embedding methods

We will begin by covering classic word embedding methods. Through a mixture of lectures and practical exercises, participants will review 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 introduce GloVe and word2vec – popular word embedding methods – and illustrate their commonalities and differences.

In the afternoon, we will present examples and exercises how to use pre-trained word embedding models to perform basic computations such as finding similar terms, computing words' similarities, and inducing conceptual dimensions.

During the exercise time, the instructors will be available for troubleshooting technical issues participants encountered when trying to setup 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: Contextualized embeddings, attention, and Transformers

Building on the topics covered on the first day, we will discuss the limitations of classic 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.

Building on this intuition, the instructors will introduce Transformer models like BERT and GPT, and discuss their advantages over traditional NLP models. We delve directly into the subject matter through practical exercises illustrating how Transformer models overcome the multiple word senses problem of "traditional" word embeddings by generating contextualized word embeddings.

We will then explore how Transformer models can be used in social science research. We will begin by focusing on text classification applications. The instructors will explain and illustrate with the Hugging Face's Transformers library how pre-trained Transformer models like BERT can be fine-tuned for various classification tasks, such as classifying the sentiment of tweets, detecting mentions of social groups in political texts, and comparing pairs of texts in terms of their emotional intensity.

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. (2021). Introduction to Neural Transfer Learning with Transformers for Social Science Text Analysis. *Sociological Methods & Research*. https://journals.sagepub.com/doi/full/10.1177/00491241221134527
- 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/
- 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

Day 3: (More) Fun with Transformers

On the third day, we will continue to explore how Transformer models can be used in social science research. We will begin with a follow-up to the exercises from the second day where participants will take a close look at how to evaluate classification performance.

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 learn to apply this approach in Python.

Overall, the third day will be structured to give participants as much opportunity to apply the learned techniques to prepared examples or the data they bring to the course from their own research. This will ensure that everyone gets the hands-on experience that will enable participants to apply the methods and concepts we introduce in their research after the course.

Recommended literature:

- Bonikowski, B., Luo, Y., & Stuhler, O. (2022). Politics as Usual? Measuring Populism, Nationalism, and Authoritarianism in U.S. Presidential Campaigns (1952–2020) with Neural Language Models. *Sociological Methods & Research*, *51*(4), 1721-1787. https://doi.org/10.1177/00491241221122317
- 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*, *0*(0). https://doi.org/10.1177/00491241221134526
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXix preprint*. https://arxiv.org/abs/2203.05794
- 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

- Licht, H., & Sczepanski, R. (2024). Who are they talking about? Detecting mentions of social groups in political texts with supervised learning. *OSF preprint*. https://doi.org/10.31219/osf.io/ufb96
- Wang, Y. (2023). On Finetuning Large Language Models. *Political Analysis*, first view 1–5. https://doi.org/10.1017/pan.2023.36

Day 4: LLM prompting and in-context learning

On day four we will move fast forward to catch up with current developments in the computational text analysis literature in the social sciences by focusing on using LLMs for text analysis through "prompting" and "in-context learning." Instructed to perform a task like sentiment classification, LLMs like ChatGPT often exhibit reliable instructing-following behavior. This opens new opportunities for overhauling established approaches to text-based measurement and inventing new ones.

The instructors will begin by explaining and illustrating how conversational assistants like ChatGPT are trained to make them instruction-following. Next, they will introduce best practices for prompt writing and development.

We will then focus on implementing the text classification tasks introduced on Day 2 with ChatGPT, allowing for a head-to-head comparison between the supervised fine-tuning and LLM prompting approaches to text classification. Participants will apply these techniques with provided data or the data they bring to the course during exercises.

Recommended literature:

- 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, e2305016120. https://doi.org/10.1073/pnas.2305016120
- Burnham, M. (2023). Stance Detection With Supervised, Zero-Shot, and Few-Shot Applications. https://doi.org/10.48550/arXiv.2305.01723
- Ziems, C., Held, W., Shaikh, O., Zhang, Z., Yang, D., & Chen. J. (2023). Can Large Language Models Transform Computational Social Science? https://arxiv.org/abs/2305.03514
- 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

Day 5: Other LLM applications, ethical Considerations, and participant project 1-on-1s

We will follow up on the previous day by providing a high-level overview of tasks other than text classification LLMs can perform, such as extractive summarization and topic modeling.

We will then shift our focus to important ethical considerations of using LLMs and other advanced deep learning methods in social scientific research.

In the afternoon of day 5, the instructors will meet with participants in 1-on-1 sessions to discuss their research projects, answer questions, etc. on an individual 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