GESIS Fall Seminar in Computational Social Science 2023

Syllabus for week 3:
"From Embeddings to Transformers: Advanced Text Analysis with Python"

Lecturers:	Hauke Licht	Jennifer Victoria Scurrell
Affiliation:	University of Cologne	ETH Zurich
Email:	hauke.licht@wiso.uni-koeln.de	jennifer.scurrell@sipo.gess.ethz.ch

Date: September 25-29, 2023
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 computational text analysis methods to study political communication, electoral competition, and democratic representation. He also has a strong focus on multilingual analyses. In this research, he increasingly uses deep learning methods to analyze textual and audio-visual data.

Jennifer Victoria Scurrell is a PhD candidate at the Center for Security Studies (CSS) at ETH Zurich. In her dissertation project she examines the impact of GPT-3 based and AI-enhanced bots on political opinion formation in online social networks using empirical experimentation, agent-based modeling, and machine learning. Furthermore, she is interested in investigating communicative influence with Natural Language Processing (NLP) methods, modifying opinion dynamics models by adding real world data, and analyzing visual political communication using deep learning. Jennifer is a political scientist by training and holds a MA and a BA from the University of Zurich.

Course Description

This course introduces social scientists to advanced, deep-learning based text analysis methods such as word embedding and large neural language models such as the Transformer. Basic methods of text analysis like counting words or *n*-grams have limitations in handling the complexity of natural language. By allowing to capture the semantic relationships between words in the contexts in which they appear, text embedding methods and neural language modeling techniques help overcome these limitations.

Participants will **learn about the conceptual motivation and methodological foundations of text embedding methods and large neural language models**. 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. Specifically, participants will apply word embeddings, for example, to identify relevant keywords when expanding a dictionary or to identify semantic dimensions in their corpus such as an emotion–reason dimension. In a second part of the course, we will introduce state-of-the-art Transformer models like BERT and GPT. We will first cover their methodological foundations: the attention mechanism, (masked) language modeling, and the encoder-decoder architecture. Participants will then apply these models in exercises covering supervised learning and topic modeling with BERTopic. 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.

Keywords

Computational text analysis, Word embedding, Large language models, Transformers, Deep learning, Python

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
  - using loops and defining new functions

For those who would like a primer or refresher in Python, we recommend taking the online workshop "Introduction to Python" that takes place from 04-06 September 2023

- Basic knowledge of quantitative research methods
  - knowledge of basic statistics (distributions, correlation)
  - understanding of linear and logistic regression analysis
  - a basic understanding of matrix algebra might be helpful but is not required

Target Group

Participants will find the course useful if:

- they have a background in the social sciences or humanities (e.g., communication science, economics, political science, sociology, or related fields)
- they have a solid understanding of basic text analysis methods and want to advance their knowledge, skills, and practical experience

Course and Learning Objectives

By the end of the course participants will:

- know the methodological foundations of text embeddings methods, large neural language models (at a conceptual level)
- be able to apply these methods to analyze social science text data
- be able to reflect critically about the application of the techniques in social science research, including relevant ethical considerations

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.

Software and Hardware Requirements

- Participants should bring their own laptops.
- They should have Python (≥ 3.10), conda, and Jupyter Notebook installed (see this link)

- Required Python libraries
  - text processing: `nltk`, `scikit-learn`, `gensim`, `transformers`
  - others: `numpy`, `scipy`, `tqdm`
- 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.

Day-to-day Schedule and Literature

Required readings are marked with a star ★.

Day 1: Introduction & Overview of Word Embedding Methods

Morning sessions:

We will begin the first day by getting to know each other and use this as an opportunity to learn about everyone's motivations to participate in the course. We will then outline the day-by-day schedule of the course.

In the second half of the morning session, we will begin the first of two thematic blocks of the course 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.

Afternoon sessions:

In the afternoon session, we will introduce GloVe and word2vec – the popular word embedding models – and illustrate their commonalities and differences. However, the last part of the afternoon session will be reserved for our course-internal *Help Café*: We will ensure that everyone's Python environment and Jupyter Notebook setup is working and troubleshoot any remaining technical issues.

Literature:

- High-level overview and primer
  - ★ 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
  - ★ Arseniev-Koehler, A. (2022). Theoretical Foundations and Limits of Word Embeddings: What Types of Meaning Can They Capture? *Sociological Methods & Research*. https://doi.org/10.1177/00491241221140142
- Methodological background
  - 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
  - Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. https://doi.org/10.1613/jair.2934
  - Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781
  - Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*. URL
  - Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

- Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 238–247. https://doi.org/10.3115/v1/P14-1023

Day 2: Using, Training & Validating Word Embeddings

Morning sessions:
On the second day of the course, we will focus on using word embedding models. We will begin by demonstrating how to work with pre-trained word embedding models. Participants will then learn and practice how to compute with embeddings. For example, we will learn how to assess the similarity between two words by computing the cosine similarity between their embeddings. We will build on these examples to learn techniques for assessing the quality and validity of embeddings.

Afternoon sessions:
In the first part of the afternoon session, we will then move on and show how to *train* an embedding model "from scratch" (i.e., on a new text corpus), using a number of pre-selected text corpora taken from the domains of politics and the media. We will take this exercise as an opportunity to cover the foundations of deep learning (back propagation and stochastic gradient descent).

Commented [1]: @VERENA: Is this too techy?

Commented [KV2R1]: I think it's a good idea for participants at this level to give them some intuition of what's going on "under the hood" of the models that they will be using, especially if this provides a chance to build a bridge between computer sciency speak and social sciency concepts that they might actually be familiar with. I'm not sure how much time you have scheduled for this currently, but would advise to probably not spend more than an hour on this to not entirely lose participants who might not be able to follow and/or are not interested in the techy parts.

Commented [MOU3R1]: Alright

In the second part of the afternoon session, we will then illustrate how one can employ word embeddings in social science research. We will focus our attention on two particular techniques: using word embeddings for dictionary expansion/keyword discovery and for constructing/extracting semantic dimensions.

Literature:

- Training and validating word embeddings
  - Yin, Z., & Shen, Y. (2018). On the Dimensionality of Word Embedding. *Advances in Neural Information Processing Systems*, *31*. URL
  - ★ Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 298–307. https://doi.org/10.18653/v1/D15-1036
  - Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C.-C. J. (2019). Evaluating Word Embedding Models: Methods and Experimental Results. *APSIPA Transactions on Signal and Information Processing*, *8*, e19. https://doi.org/10.1017/ATSIP.2019.12
  - Allen, C., & Hospedales, T. (2019). *Analogies Explained: Towards Understanding Word Embeddings* (arXiv:1901.09813). arXiv. http://arxiv.org/abs/1901.09813
- Social science applications
  - Dictionary expansion/keyword discovery
    - Osnabrügge, M., Hobolt, S. B., & Rodon, T. (2021). Playing to the Gallery: Emotive Rhetoric in Parliaments. *American Political Science Review*, *115*(3), 885–899. https://doi.org/10.1017/S0003055421000356
    - 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
  - Extracting conceptual dimensions
    - ★ 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

Day 3: Applications, Limitations & Extensions of Word Embedding Models

gesis training

gesis Leibniz-Institut für Sozialwissenschaften

Morning sessions:
In the morning session, we will continue the block from the previous day by introducing another important use case of word embeddings: to create document representations that can be used in downstream analyses such as supervised classification.

We will continue the morning session by pointing participants to two important advanced uses and extensions of standard word embedding models in social science research: (i) measuring over-time shifts in word meaning using dynamic embedding methods and (ii) computing embeddings for documents. Last but not least, 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.

Afternoon sessions:
In the afternoon of day 3, we move on to the second thematic block focusing on transformer models. The instructors provide a brief introduction to transformer models and 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.

Literature:

- Using word embeddings for supervised classification
    - Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, *12*(2–3), 140–157. https://doi.org/10.1080/19312458.2018.1455817
- Using word embeddings for measuring/discovering over-time semantic shifts
    - Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115
    - 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
- Document embeddings
    - Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *International Conference on Machine Learning*, 1188–1196. http://proceedings.mlr.press/v32/le14.html
    - Arora, S., Liang, Y., & Ma, T. (2019). *A simple but tough-to-beat baseline for sentence embeddings*. 5th International Conference on Learning Representations, ICLR 2017.
    - 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
- Transformer models and contextualized embeddings
    - Liu, Q., Kusner, M. J., & Blunsom, P. (2020). A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*. 2003.07278.pdf (arxiv.org)
    - ★ Smith, N. A. (2020). Contextual word representations: putting words into computers. *Communications of the ACM*, *63*(6), 66-74. https://dl.acm.org/doi/pdf/10.1145/3347145

Day 4: Understanding & Applying Transformers and BERT

Morning sessions:
During the morning session, we explore how transformer models can be used in social science research. Sentiment analysis, fake news detection, and topic identification are all high-level NLP tasks that can be accomplished with cutting-edge transformer models. We will discuss how model pre-training and fine-tuning work. We will deepen participants' understanding of neural language modeling by doing exercises on masked language models.

Afternoon sessions:

In the afternoon, participants will learn in a series of hands-on exercises how to fine-tune transformer models for different NLP tasks, such as supervised text classification, with the Hugging Face's `transformers` library.

Literature:

- High-level overview and primer
  - ★ 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
- Methodological foundations
  - 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://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
  - Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6795963
- BERT
  - 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/
  - Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. https://arxiv.org/ftp/arxiv/papers/2103/2103.11943.pdf

Day 5: Advanced BERT Applications, Other Large Language Models & Ethical Considerations

Morning sessions:
In the morning of day 5, the instructors will introduce BERTopic – a BERT-based approach to topic modeling. Participants will have time to experiment with this method through hands-on exercises.
We will then shift our focus to large language models like GPT. Participants are introduced to ChatGPT and we discuss ethical considerations of large language models.

Afternoon sessions:
In the afternoon of day 5, we recapitulate the course material of the previous days and answer open questions. At the end of the session, there will be time for 1-on-1 meetings where participants can consult the instructors with questions and problems they face with their research projects.

Literature:

- Topic modeling with BERT
  - Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv*:2203.05794. https://arxiv.org/abs/2203.05794
- GPT
  - Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, 1(6), 831-833. https://reader.elsevier.com/reader/sd/pii/S2667325821002193?token=3B02A242663562D5A19CCBB730E7E20CCD8DC5182DE40F244BB139BDFB0DD07C611B4B158130765AE823A1F5A1243D83&originRegion=eu-west-1&originCreation=20230328121432
- Risks and ethics
  - 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. 2108.07258.pdf (arxiv.org)
  - ★ 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

- 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 arXiv:2301.04246*. 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

Additional Recommended Literature

fast.ai's free *Practical Deep Learning for Coders* online course: https://course.fast.ai/

Raschka, S., Liu, Y., Mirjalili, V., & Dzhulgakov, D. (2022). *Machine learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python*. Packt Publishing. Published online

Tunstall, L., Werra, L. von, Wolf, T., & Géron, A. (2022). *Natural language processing with transformers: Building language applications with hugging face* (Revised edition). O'Reilly. URL

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