Discipline: special methods

1. Language

English

2. Title

Natural Language Processing in Business Research

3. Lecturer

Dr. Matthias Aßenmacher, Department of Statistics, LMU Munich https://www.slds.stat.uni-muenchen.de/people/assenmacher/ matthias@stat.uni-muenchen.de

Dr. Jan Klostermann, University of Cologne
https://marketing.uni-koeln.de/en/team/jan-klostermann klostermann@wiso.uni-koeln.de

4. Date and Location

Date: 21.09.2026 – 24.09.2026

Location: Cologne

The course will be offered over four days and will comprise lectures, tutorials, and discussion sessions.

5. Course Description

5.1 Learning Objectives

Customers and employees write millions of reviews every day, companies publish multiple reports every year, and important stakeholders publicly debate on social media all the time. This course teaches business researchers how to leverage these vast amounts of textual data in their research.

This graduate-level course provides a rigorous and self-contained treatment of deep learning (DL) for natural language processing (NLP), focusing on the architectures and methods driving current research on large language models (LLMs). While foundational concepts, such as tokens and embeddings, are covered to ensure a shared baseline, the course quickly advances to recent topics including encoder-based models (BERT and beyond), multitask learning, and generative LLMs.

The course is structured across three dimensions: methodological depth, hands-on coding, and applied research. Students develop a thorough understanding of key NLP architectures, gain practical programming skills to fine-tune models and work with generative AI tools, and explore

how these methods can be integrated into business and economics research through dedicated application sessions and research presentations. A central goal is to equip graduate students not only to understand current NLP methods, but to independently apply and critically reflect on them within their own research.

5.2 Content

The course provides a comprehensive overview of the state-of-the-art (SOTA) ML and DL approaches for using unstructured text in applications in business and economics. The course can be roughly categorized into four parts.

Part I introduces the usage of text data in business research in general and showcases the numerous opportunities as well as challenges arising from using such data in their own research. The remainder of this part introduces foundational concepts central to contemporary NLP, like RNNs, the attention mechanism, (ELMo) embeddings, and tokenization. Each concept serves as a building block in understanding how neural networks can comprehend and generate human language. Subsequently, we introduce the Transformer (Vaswani et al., 2017), a DL architecture specifically designed for sequence-to-sequence tasks in natural language processing. It revolutionized NLP by replacing recurrent layers with self-attention mechanisms, enabling it to process entire sequences in parallel, overcoming the limitations of sequential processing in traditional models. This architecture has become the foundation for state-of-the-art models in NLP, Vision, or sequence processing tasks in general, efficiently solving tasks such as machine translation, text summarization, and language understanding.

In **Part II**, we explore different parts of the Transformer (Encoder and Decoder), and discuss SOTA transformer-based models before focusing on their practical implementation. For encoder-based models we focus on BERT (Develin et al., 2019) and discuss the concepts of pre-training, fine-tuning, transfer learning, and self-supervision. We further discuss T5 (Raffel et al., 2020) as a representative for encoder-decoder-based models and go in-depth on concepts central to this setup, including text-to-text formulation of tasks, multi-task learning, and cross-task transfer.

Part III covers topic modeling, a technique for discovering hidden themes in text data. We focus on Latent Dirichlet Allocation (LDA), a probabilistic model that assigns words to topics based on co-occurrence patterns. We explore important extensions such as dynamic, guided, and supervised topic models that help to understand how topics evolve and which topics drive relevant outcomes (e.g., the number of stars in online reviews). Further, we learn how deep learning-based methods such as Topic-BERT can utilize contextual embeddings for more coherent topic discovery. We further discuss evaluation metrics and practical applications of topic modeling and learn how to implement these models on real datasets from marketing, accounting, and human resources.

In **Part IV**, we will finally discuss generative LLMs based on the transformer-decoder. Students will first learn about the evolution of the GPT series, spanning from GPT-1 to GPT-3, which revolutionized natural language processing by employing generative transformer architectures pre-trained on massive text corpora to generate contextually relevant text. We will also continuously discuss how LLMs are used in business research (e.g., to analyze customers), ethical implications, and potential societal impact, including issues surrounding bias, misinformation, and data privacy. Building upon this foundation, we cover important LLM concepts, such as Instruction Fine-Tuning, Chain-of-Thought prompting, and Reinforcement Learning from Human Feedback (RLHF). Before SOTA LLMs, such as Llama, Phi, or the Mistral models, are introduced, students will be taught different decoding strategies for using

generative LLMs as well as how to consider the massive implications such models have on the computational side.

5.3 Course Schedule

The course comprises lectures for methods (•), programming (•), and applied research (•) sessions.

Pre-course stage		
		(Study papers from the reading list) Familiarize yourself with Python and Jupyter notebooks
Day 1 - Introduction		
		Arrival of participants
11:00	12:00	General introduction and motivation •
12:00	13:00	Lunch
13:00	15:00	Basics: Tokens, Embeddings, RNNs, Attention, NLP Tasks •
15:00	17:00	Foundation: The Transformer inside out •
Day 2 - Transformers and Fine-Tuned Models		
09:00	11:00	Encoder-Based Models: BERT (and Colleagues) •
11:00	12:00	Fine-Tuning Encoder-Based Models for Classification •
12:00	13:00	Lunch break
13:00	15:30	Research Presentations: Challenges and opportunities (Assignment) •
15:30	17:00	Multitask Learning and Text-to-Text Task Formulation •
Day 3 – Large language models		
09:00	11:00	Topic Modeling •
11:00	12:00	Applied topic modelling in business research •
12:00	13:00	Lunch break
13:00	15:00	Generative LLMs I •
15:00	17:00	Generative LLMs II •
Day 4 - Generative Models		
09:00	11:00	Applications of Generative LLMs in business research •
11:00	12:00	Using Generative LLMs in business research •
12:00	13:00	Lunch break
13:00	16:00	Research Presentations: Reflect ideas for solutions (Assignment) •
16:00	16:30	Closing: Final remarks and Q&A
Post-course stage (Exam)		
4 to 6 weeks		Development of a programming script (e.g., Jupyter notebook) demonstrating the use of NLP for practical problem-solving in business and economics research. Specific tasks will be agreed upon with participants and should ideally display a strong link to the participant's Ph.D. topic.

5.4 Course format

The course adopts a multi-faceted teaching concept combining conceptual lectures, discussion, reviews of programming codes, and hands-on exercises using Python. Each of the three core parts is associated with programming demos and exercises using real-world data sets from fields such as marketing, management, finance, and human resources. The data will be provided in the course but participants are encouraged to bring their own dataset.

The final assignment will allow students to deepen their practical skills by working on an NLP task, which can be connected and/or beneficial to their Ph.D. research.
The course language is English.

6. Preparation and Literature

6.1 Prerequisites

Master-level education in Business, Economics, Computer Science, Engineering, or a related field.

Course participants will benefit from some basic knowledge of Python programming. Practical exercises and assignments will use the Python programming language. Therefore, familiarity with Python, virtual environments, and Jupyter Notebooks is beneficial, but can also be acquired in the pre-course stage.

Prior knowledge from related VHB-ProDok courses such as Data Science as a Research Method (Oliver Müller) or Machine Learning (Stefan Lessmann) is beneficial, though not required.

6.2 Essential Reading Material (larger books, read/skim selected chapters)

- *Setting the narrative from a business perspective:*
  Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. Journal of Marketing, 84(1), 1-25. https://doi.org/10.1177/0022242919873106
- *Non-technical Introduction to Language Processing:*
  Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., ... & Van Bavel, J. J. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, 1-16.
- *General Introduction to Language Processing (Chapters 1-5):*
  Martin, J. H., & Jurafsky, D. (2009). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (Vol. 23). Upper Saddle River: Pearson/Prentice Hall. https://web.stanford.edu/~jurafsky/slp3/ed3book_Jan25.pdf
- *Overview on Embeddings (Chapters 1-3):*
  Pilehvar, M. T., & Camacho-Collados, J. (2020). Embeddings in natural language processing: Theory and advances in vector representations of meaning. Morgan & Claypool Publishers. https://doi.org/10.1007/978-3-031-02177-0
- *NLP with neural networks (Chapters 6-9):*
  Goldberg, Y. (2017). Neural network methods in natural language processing. Morgan & Claypool Publishers. https://doi.org/10.1007/978-3-031-02165-7

6.3 Additional Reading Material (Specific Papers)

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022. https://dl.acm.org/doi/10.5555/944919.944937
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In The First International Conference on Learning Representations (ICLR), Scottsdale, AZ, USA, May 2-4, 2013, Conference Track Proceedings. https://doi.org/10.48550/arXiv.1301.3781
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA (pp. 6000-6010). https://doi.org/10.48550/arXiv.1706.03762

6.4 To prepare

Participants are expected to study the essential reading material. Familiarity with literature from the additional reading material is not required but beneficial. To prepare for the practical exercises and course assignments, participants are required to familiarize themselves with the Python programming language and Jupyter notebooks. To that end, participants might find the following textbook useful:

- Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd Edition. O'Reilly Media Inc.
- VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. Sebastopol, CA, USA: O'Reilly Media. https://jakevdp.github.io/PythonDataScienceHandbook/

Participants are expected to prepare a presentation about their current research with a focus on challenges and opportunities related to NLP. If their current research has no connection to NLP, a short paper presentation can be prepared as a substitute. Details about this presentation will be shared in the month before the course.

7. Administration

7.1 Max. number of participants

The number of participants is limited to 20.

7.2 Assignments

Participants who seek to earn credits have to (1) give two research presentation and are expected to (2) actively participate in all in-class discussion.

7.3 Exam

After the course, participants are required to complete a NLP assignment and write-up results in the form of a computational essay (i.e., Jupyter Notebook). Ideally, the assignment task connects to a research project that the participant is involved. The schedule of the course leaves room for discussing possible topics for the assignment. Student will submit their solution to the assignment roughly six weeks after the end of the course period. The submitted notebooks will be graded according to the quality of the exposition, the complexity of the modeling tasks, and the degree to which NLP concepts have been used successfully.

7.4 Credits

The course corresponds to a scope of 6 LP/ECTS

8. Working Hours

Working Hours	Stunden
Preparation (reading material, software, presentation)	70h
Active participation	30h
Exam (Post-course stage)	80h
SUMME	180 h