GESIS Fall Seminar in Computational Social Science 2025

"Introduction to Machine Learning for Text Analysis with Python"

Lecturers:	Damian Trilling	Rupert Kiddle
Affiliation:	Vrije University Amsterdam	Vrije University Amsterdam
Email:	d.c.trilling@vu.nl	r.t.kiddle@vu.nl

Date: September 15-19, 2025
Time: 10:00-17:00

About the Lecturers

Damian Trilling is a full professor at Vrije Universiteit Amsterdam and holds the Chair for Journalism Studies. He is interested in transformations of news environments and the question how users engage with and are exposed to societally relevant information. To do so, he uses computational methods, in particular for text analysis. He is co-author of a textbook on the computational analysis of textual data and has been teaching Python to social scientists for more than a decade.

Rupert Kiddle is a Doctoral Candidate in Computational Communication Science at the Department of Language, Literature, and Communication, Vrije University Amsterdam. His research leverages computational methods to reveal patterns and dynamics of online news consumption via digital trace data. Additionally, it examines the effectiveness of language modeling in developing content-based news recommendation architecture that fosters serendipity, encouraging users to explore and discover novel and diverse journalistic content.

Course Description

The course will provide insights into the concepts, challenges and opportunities associated with data so large that traditional research methods (like manual content analysis) can no longer be applied, and traditional inferential statistics start to lose their meaning. Participants are introduced to strategies and techniques for capturing and analyzing digital data in communication contexts using Python. The course offers hands-on instructions regarding the several stages of computer-aided content analysis. More precisely, students will be familiarized with pre-processing methods, analysis strategies and the visualization and presentation of findings. The focus will be in particular on Machine Learning techniques to analyze quantitative textual data, amongst which both deductive (e.g., supervised machine learning) and inductive (e.g., unsupervised machine learning) approaches will be discussed.

This is a beginner's course. Participants who are looking to learn about the latest developments in machine learning for textual data (such as large language models) should consider taking a different course, e.g., "From Embeddings to LLMs: Advanced Text Analysis in Python" (22-26 September). These techniques will be (briefly) discussed towards the end of the course, but the focus lies on the basics of natural language processing and classical machine learning in Python.

Organizational Structure of the Course

In the morning, we will have lectures, in which we will explain the topic of the day both from a theoretical-conceptual point of view as well as from a practical point of view (i.e., walking you through code examples). We may have small in-class exercises in between, if necessary.

In the afternoon, students work on larger exercises in which they implement the techniques we covered. We provide example datasets, but it is also possible (and strongly encouraged) to try to apply the techniques to own datasets. Participants can either opt to work on their own or try to solve problems together with one or multiple classmates. Lecturers will provide feedback on the (attempted) solutions of participants and also provide example solutions.

Keywords

Language Modelling, Supervised Machine Learning, Unsupervised Machine Learning

Target Group

You will find the course useful if:

- You are a social scientist who has the ambition to model quantitative textual data. Specifically, those who aim to describe, explain, or predict the content of large-scale textual data using computation techniques are likely to benefit from participating in this course.
- Note that non-textual data, such as images or networks, are not at the center of this course. Techniques we cover are partly generalizable to such types of data but the course is not tailored towards them. If you want to work with images or networks, you might be interested in one of the following courses: "Computer Vision for Image and Video Data Analysis with Python" (15-19 September) or "Advanced Methods for Social Network Analysis" (15-19 September).

Course and Learning Objectives

By the end of the course, you will:

- be able to identify research methods from computer science and computational linguistics which can be used for research in the domain of social science
- have an understanding of the principles of supervised and unsupervised machine learning
- be able to explain the principles of these methods and describe the value of these methods
- know how to analyze textual data
- have basic knowledge of the programming language Python and know how to use Python-modules for questions relevant in the domain of the social sciences
- be able to independently analyze quantitative textual data using machine learning techniques

Course Prerequisites

- Knowledge of basic statistics (linear and logistic regression)
- Some experience with computational methods, programming in general, and/or statistical languages (but not necessarily Python) is highly recommended to participate in this course. During the first day of the course, we will discuss some fundamental aspects of coding in Python at a fast pace. In order to follow along, we recommend those who have little previous experience with computational methods or statistical languages to take part in the online blended learning course "Introduction to Computational Social Science with Python" (01-05 September).
- Participants are expected to have a working Python environment installed (see below), and we strongly recommend that participants spend a couple of hours with one of the many free online resources to familiarize themselves with the very basics of Python to have an easier start. For a basic introduction or refresher to Python programming, participants may also consider taking the online workshop "Introduction to Python" that takes place from 25-28 August.

Software and Hardware Requirements

Participants should bring their own laptops for use in the course. You need to have a current Python environment installed and need to be able to install and update packages on your own. In the weeks prior to the course, the lecturers will distribute a setup guide which will ensure that participants can work within a similar environment to the lecturers. This guide is entirely optional, as you may alternatively set up your environment to your own preferences. All relatively recent versions of Python (in general, 3.8 or higher) should be fine. If you still have an older version, you may not be able to run the example code 1:1 but need to adapt it. Make sure you have recent versions of crucial packages such as pandas, numpy, scipy, scikit-learn, gensim, keras, sentence transformers and BERTopic installed. Additionally, it is advisable to have access to Google Colab and (Microsoft) GitHub. Therefore, please ensure that you have Google and Microsoft accounts, and can execute code through Google Colab.

Course Contents

- Python
- Language Modelling
- Unsupervised Machine Learning
- Supervised Machine Learning

Day-to-day Schedule and Literature

Day 1: Introduction and Getting Started in Python

- Introduction and overview of the course
- Principles of quantitative textual analysis for social scientists
- Getting started with programming in Python: Introduction to the main concepts (such as data types, functions, and methods)
- Practical discussion of benefits and drawbacks of working with different IDEs, as well as working with specific modules (such as pandas) versus native Python data structures
- Conducting an exercise that focuses on setting up our first simple machine learning classifier

Suggested reading:

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 1–12. doi:10.1177/2053951714528481

van Atteveldt, W., Trilling, D., Arcila Calderon, C. (2022): *Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R.* Hoboken, NJ: Wiley. *Chapters 1-4.* https://cssbook.net/

Day 2: Preparing for Analysis: From Text to Features

- Introduction to the toolkit accessible to social scientists working with 'big' textual datasets
- Inductive and deductive approaches to computer-aided content analysis
- Exploratory techniques to explore your data
- When, why, and how do we pre-process?
- Regular expressions and their application
- Natural Language Processing with NLTK and spacy
- From text to features: count vectorizers and tf-idf vectorizers

Suggested reading:

Boumans, J. W. & Trilling, D. (2016). Taking stock of the toolkit. An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism, 1*, 8-23. doi:10.1080/21670811.2015.1096598

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. Information Systems, 121, 102342. doi:10.1016/j.is.2023.102342

van Atteveldt, W., Trilling, D., Arcila Calderon, C. (2022): *Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R.* Hoboken, NJ: Wiley. *Chapters 9-10.* https://cssbook.net/

Day 3: Unsupervised Machine Learning

- Principles and techniques of Unsupervised Machine Learning techniques
  - e.g., a brief introduction to Principal Component Analysis, k-means clustering, and hierarchical clustering
- Exercises to apply these techniques, using modules such as scikit-learn and gensim
- Hands-on topic modeling with transformers: *BERTopic*.
- Comparing techniques of unsupervised learning with supervised learning

Suggested reading:
Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv. https://arxiv.org/abs/2203.05794
van Atteveldt, W., Trilling, D., Arcila Calderon, C. (2022): *Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R.* Hoboken, NJ: Wiley. *Chapters 7.3 and 11.5.* https://cssbook.net/

Day 4: Supervised Machine Learning

- Principles and techniques of Supervised Machine Learning
- Discussion of how logistic regression and Naive Bayes classifiers can be used to predict, for instance, movie ratings or topics of news articles
- Understanding evaluation metrics (accuracy, precision, recall, ...)
- Hands-on instructions to apply these techniques, using modules such as scikit-learn
- Alternative models (e.g., Random Forests)
- Advanced Supervised Machine Learning (e.g., cross-validation, grid search, model selection, and tuning)

Suggested reading:
van Atteveldt, W., Trilling, D., Arcila Calderon, C. (2022): *Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R.* Hoboken, NJ: Wiley. *Chapters 8 (except 8.4) and 1.* https://cssbook.net/

Day 5: Recent Developments in Machine Learning

- Discussion of Transfer Learning and Transformers
- Training and Fine-Tuning BERT for classification (supervised)
- Introduction to Large Language Models (LLM's) for classification
- Navigating the landscape: Making sense of state-of-the-art methods, their trade-offs in performance, computational demands, and accessibility for social science application

Suggested reading:
Kroon, A., Welbers, K., Trilling, D., & van Atteveldt, W. (2024). Advancing automated content analysis for a new era of media effects research: The key role of transfer learning. Communication Methods and Measures, 18(2), 142-162. doi: 10.1080/19312458.2023.2261372
van Atteveldt, W., Trilling, D., Arcila Calderon, C. (2022): *Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R.* Hoboken, NJ: Wiley. *Chapter 8.5.* https://cssbook.net/