

## Introduction, administrivia

Statistical Methods in NLP 1  
ISCL-BA-06

Çağrı Çöltekin  
/çac@uni-tuebingen.de  
ccoltekin@fsa.uni-tuebingen.de

University of Tübingen  
Seminar für Sprachwissenschaft

Winter Semester 2025-2026

version: 03/08/2022-02-03

## What is this course about?

- This is the first course in the two-course series on *Natural Language Processing*
- A course on foundations of machine learning
- Focus on theory and concepts (a bit of practice in the second part)

### Prerequisites:

- Formally: ASW-BA-01, ISCL-BA-01, ISCL-BA-04
- Realistically: High-school math and some programming (in Python)

Module: ISCL-BA-06 (formerly called Parsing)

Ç. Çöltekin, SSF | University of Tübingen

Winter Semester 2025-2026 1 / 1

## What is NLP?

NLP is a set of methods for computational processing of natural languages.

- Is it a subfield of:
  - Artificial Intelligence?
  - Machine Learning?
  - Computer Science?
  - Information Retrieval? – Wikipedia claims so.
- Is it the same as Computational Linguistics?
- How much linguistics is needed / used?

Ç. Çöltekin, SSF | University of Tübingen

Winter Semester 2025-2026 2 / 1

## Text (sequence) classification

- Text classification is a very common NLP task
- Given a text we often want to assign one or more labels
- Assignment of the label(s) requires some level of 'language understanding'

Ç. Çöltekin, SSF | University of Tübingen

Winter Semester 2025-2026 3 / 1

## Text classification: some examples

is it spam?

From: Dr Pius Ayim <mikehass15@gmail.com>  
Subject: Dear Friend / Lets work together  
Dear Friend,  
My name is Dr. Pius Ayim, former senate president of the Republic Nigeria under regime of Jonathan Good-luck.  
I am sorry to invade your privacy; but the ongoing ANTI-CORRUPTION GRAFT agenda of the ruling government is a BIG problem that I had to get your contact via a generic search on internet as a result of looking for a reliable person that will help me to retrieve funds I deposited at a financial institute in Europe.  
-

\* From my 'spambox' which I stopped checking regularly long time ago.

Ç. Çöltekin, SSF | University of Tübingen

Winter Semester 2025-2026 4 / 1

## Text classification: some examples

is the customer happy?

I never understood what's the BIG deal behind this album. Yes, the production is wonderful but the songwriting is childish and rubbish. They definitely can not write great lyrics like Bob Dylan sometimes do. "God Only Know" and "Wouldnt Be nice" are indeed masterpieces...but the rest of the album is background music.

€85\_Bahn müßten sie für den Sauna-Besuch zurahlen ?

- Sentiment analysis is one of the popular applications of text classification

Ç. Çöltekin, SSF | University of Tübingen

Winter Semester 2025-2026 5 / 1

## Text classification: some examples

which language is this text in?

Član 3. Svako ima pravo na život, slobodu i ličnu bezbjednost.

- Detecting language of the text is often the first step for many NLP applications.
- Easy for the most part, but tricky for
  - closely related languages
  - text with code-switching

Ç. Çöltekin, SSF | University of Tübingen

Winter Semester 2025-2026 6 / 1

## Text classification: More examples

- Who wrote the book?
  - age
  - gender
  - political party affiliation
  - native language
- Is the author / speaker depressed?
- What is the proficiency level of a language learner?
- What grade should a student essay get?
- What is the diagnosis, given a doctor's report?
- What category should a product be listed based on its description?
- What is the genre of the book?
- Which department should answer the support email?
- Is this news about
  - politics
  - sports
  - travel
  - economy
- Is the web site an institutional or personal web page?

Ç. Çöltekin, SSF | University of Tübingen

Winter Semester 2025-2026 7 / 1

## Sequence labeling

Example: entity recognition

UN Secretary-General Antonio Guterres  
(ORG) (NONE) (PER) (PER)  
plans to visit Ukraine  
(NONE) (NONE) (NONE) (GEO)

- Typical entities of interest include: people, organizations, locations
- Can be application specific, e.g., drug/disease names, chemical components, legal entities
- Many NLP problems can be cast as sequence labeling problems: extractive summarization, question answering, ...

Ç. Çöltekin, SSF | University of Tübingen

Winter Semester 2025-2026 8 / 1

## Relation extraction

head-of

UN Secretary-General Antonio Guterres  
(ORG) (NONE) (PER) (PER)  
plans to visit Ukraine  
(NONE) (NONE) (NONE) (GEO)

- For many other tasks, we do not only need entities, but the relations between them
- Other similar applications include: dependency parsing, semantic role labeling, ...

Ç. Çöltekin, SSF | University of Tübingen

Winter Semester 2025-2026 9 / 1

## Text generation

Example: machine translation



Ç. Çöltekin, SSF | University of Tübingen

Winter Semester 2025-2026 10 / 1

## Text generation

Other examples

- Dialogue systems / 'intelligent' agents
- Question answering
- Summarization
- Caption generation
- Data to text generation

Models capable of generating text (LLMs) is where most of the research is focused currently.

Ç. Çöltekin, SSF | University of Tübingen

Winter Semester 2025-2026 11 / 1

## Why is it challenging

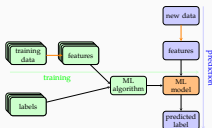
- NLP deals with combinatorial problems: high computational complexity
- Natural language is categorical on the surface, but with many nuanced/continuous differences underneath
- Linguistic input is variable length, most (if not all) ML methods require fixed-length input
- Data sparsity: many linguistic objects (e.g., words) occur rarely
- Natural language is ambiguous

## Some examples of ambiguity

(newspaper headlines)

- FARMER BILL DIES IN HOUSE
- TEACHER STRIKES IDLE KIDS
- SQUAD HELPS DOG BITE VICTIM
- BAN ON NUDE DANCING ON GOVERNOR'S DESK
- PROSTITUTES APPEAL TO POPE
- KIDS MAKE NUTRITIOUS SNACKS
- DRUNK GETS NINE MONTHS IN VIOLIN CASE
- MINERS REFUSE TO WORK AFTER DEATH

## Anatomy of a (supervised) data-driven solution



## What is in this course?

A bird's eye view

Introductory lectures on

- Background
  - Linear algebra
  - Calculus (mainly derivatives)
  - Probability and information theory
- Generalization, bias, variance
- Some fundamental classification methods
- Unsupervised learning

## Course overview

- Lectures (VG 0.02)
  - Tuesday 16:15-17:45
  - Thursday 10:15-11:45
- Tutorial (Frid: Aida Rostami)
  - Friday 10:15-11:45 (Neue Aula 05)
- Public course website: <https://anlp1-2025.github.io/>
- Moodle:  
<https://moodle.zdv.uni-tuebingen.de/course/view.php?id=1143>
- GitHub: <https://github.com/anlp1-2025/anlp1>

## Literature

- No textbook
- Reading recommendations will be provided for each course (all online, freely accessible)

## Coursework and evaluation

- Reading material for most lectures
- Assignments: ungraded, but **required**
- Final (written) exam
- Attendance is not required, but you are unlikely to pass without regular attendance

## Assignments

- Some paper & pencil assignments (through Moodle)
- 3 programming assignments (after semester break, through GitHub)
- The programming assignments can be done in pairs (recommended – knowing your classmates, and learning from them, is an important part of the university experience/education)
- This means **working together on the whole exercise**, not sharing parts of an assignment and working on them independently

## Final remarks

- Please do not be shy, ask your questions during the lectures
- Please take the assignments seriously, learning programming requires practice
- Please fill in the 'beginning of semester survey' on Moodle:  
<https://moodle.zdv.uni-tuebingen.de/mod/feedback/view.php?id=71920>
- Next:
  - Linear algebra introduction
  - Suggested **watching** video lectures:  
<https://www.youtube.com/playlist?list=PL3-GT3cukr2y2fErxmaJu2Lc94Ez005B>
- Time for your questions

## Acknowledgments, credits, references

\*