

Customizable Anonymization of German Legal Court Rulings using Domain-specific Named Entity Recognition

Master Thesis - Kickoff

Tom Schamberger, Kickoff 3rd of May 2021

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
www.matthes.in.tum.de

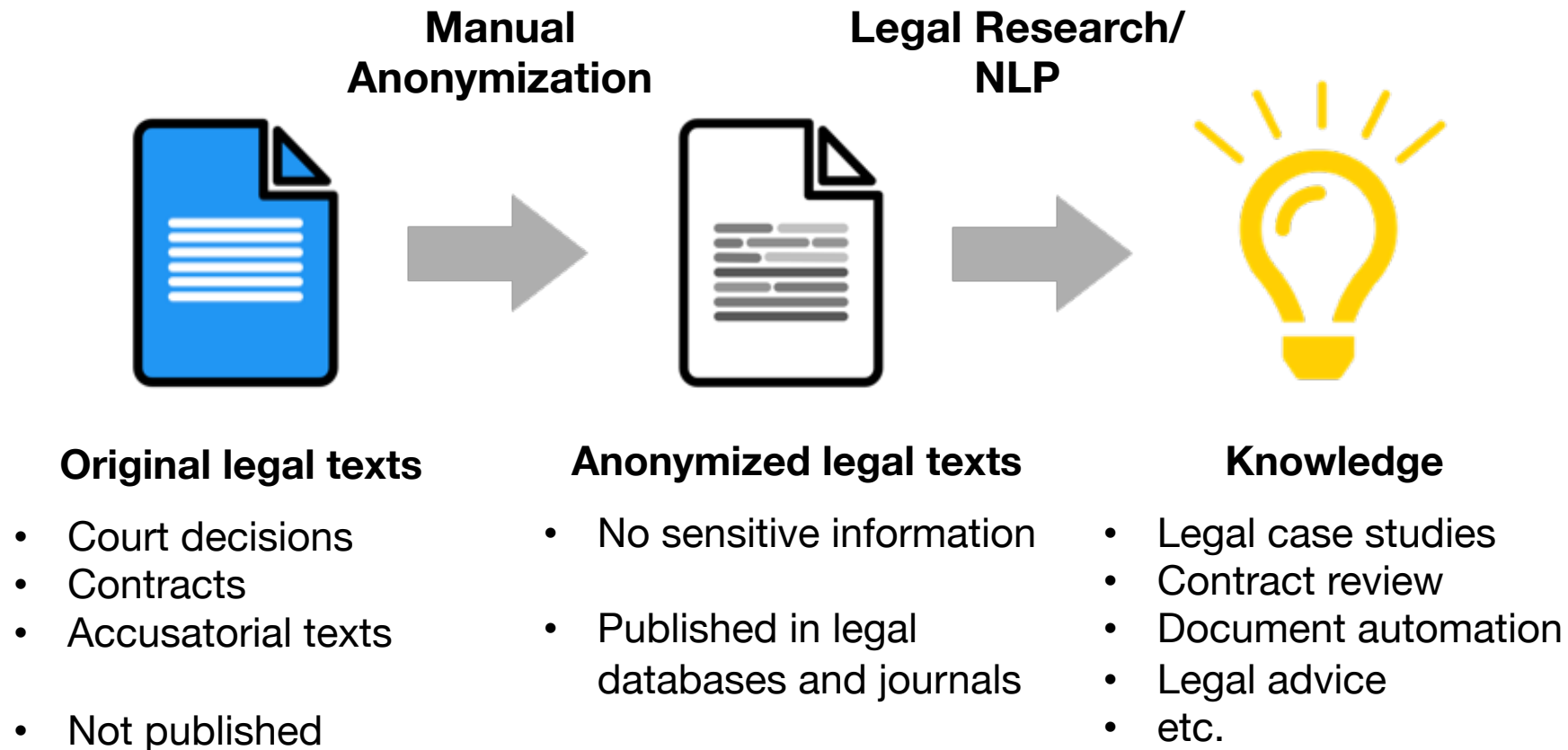
Motivation

Problem Statement

Approach

Research Questions

Schedule



- Challenges for courts
 - Mainly manual and time-consuming process
 - Prevents the publication of large amounts of court decisions
- Challenges for the processing of anonymized decisions
 - Few legal data publicly available
 - Heterogeneous anonymization conventions within and between courts
 - Anonymized entities marked using ambiguous patterns hardly detectable by machines (and even humans)

Motivation

Problem Statement

Approach

Research Questions

Schedule

Classical Automated Anonymization Approach

- Anonymization of named entities
- Usually using rule-based or dictionary-based models

Der am 16. Juni 1985 in Syrien geborene Antragsteller, nach eigenen Angaben algerischer Staatsangehöriger, reiste am 27. August 2019 auf dem Landweg in die Bundesrepublik Deutschland ein und stellte am 30. August 2019 einen Asylantrag.

Classical Automated Anonymization Approach

- Anonymization of named entities
- Usually using rule-based or dictionary-based models

Der am ... in ... *geborene* Antragsteller, nach eigenen Angaben algerischer Staatsangehöriger, reiste am ... auf dem Landweg in die ... ein und stellte am ... einen Asylantrag.

Problems:

- All named entities are anonymized, even though not all entities are sensitive
- Important information is lost

Initial approach:

Reproduce whole anonymization process through machine learning algorithm

Classification task

Der am **16. Juni** 1985 in **Syrien** *geborene* Antragsteller, nach eigenen Angaben algerischer Staatsangehöriger, reiste am 27. August 2019 auf dem Landweg in die Bundesrepublik Deutschland ein und stellte am 30. August 2019 einen Asylantrag.

Initial approach:

Reproduce whole anonymization process through machine learning algorithm

Challenges

- **Not enough original** court decisions available
 - Model training solely on anonymized data
 - Use of named existing named entity recognizers
- **Very heterogeneous data** between and within courts
- Courts wish to **customize anonymization** to meet special conditions (e.g. refugee crisis)

Example for Anonymized Court Decisions

Excerpt of two **court decisions**:

Der am 1985 in ... geborene Antragsteller, nach eigenen Angaben algerischer Staatsangehöriger, reiste am 27. August 2019 auf dem Landweg in die Bundesrepublik Deutschland ein und stellte am 30. August 2019 einen Asylantrag.

(Source: Bayern.Recht)

In court rulings concerning **asylum process**, **years of birth** are never anonymized

Die Antragstellerin ist türkische Staatsangehörige. Sie hat aufgrund einer in der Türkei am 20.02.2013 rechtskräftig gewordener Gerichtsentscheidung die beiden Kinder ..., geboren am ... in ... und ... geboren am ... in ..., adoptiert.

(Source: Bayern.Recht)

In court rulings concerning **adoption**, **years of birth** are always anonymized

Motivation

Problem Statement

Approach

Research Questions

Schedule

General NE types

Person

Date

Organization

Location

Miscellaneous

Domain-specific NE types

- Natural party (e.g. plaintiff, applicant, defendant)
- Judge
- Other persons (e.g. lawyer, expert)

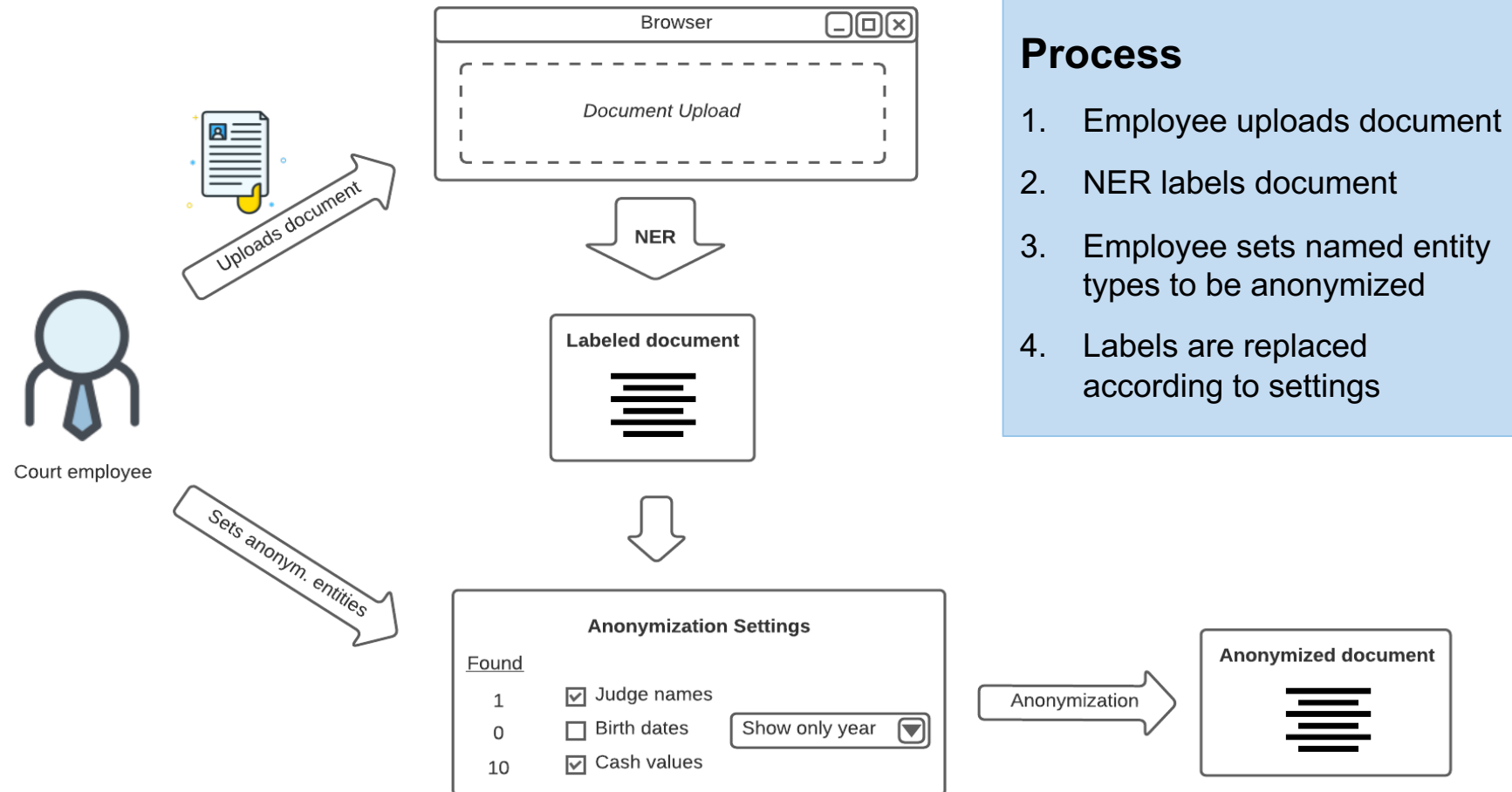
- Birth date of natural party
- Decision date
- Other dates

- Private organization
- Public organization

- Birth location / Origin of natural party
- Location of concern (of e.g. regulation, application)
- Other locations

- Brand
- Cash value
- Other

Anonymization process



1. Rewrite and label dataset of anonymized and original courts decision sections
2. Train NER model on the dataset
3. Evaluate NER model performance
4. Integrate NER into an anonymization software tool

Step 1: Dataset Preparation - Example

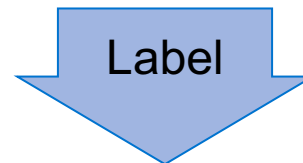
Anonymized Paragraph

Den Angeboten auf der Webseite des Klägers ... liegt das Konzept der Produktdetailseite.



"Non-anonymized" Paragraph

Den Angeboten auf der Webseite des Klägers Bayer liegt das Konzept der Produktdetailseite.



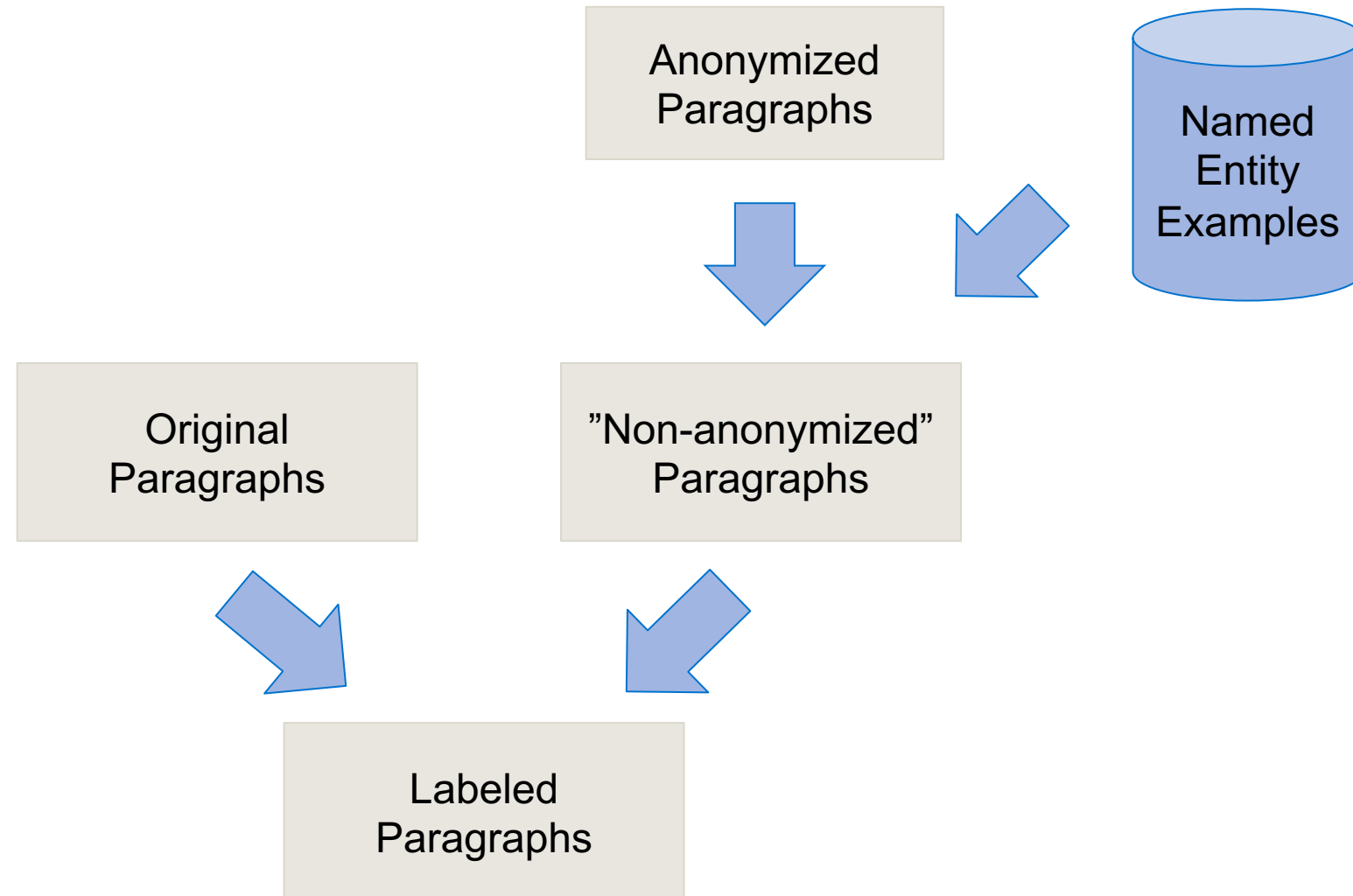
Labeled Paragraph

Den Angeboten auf der Webseite des Klägers Bayer liegt das Konzept der Produktdetailseite.



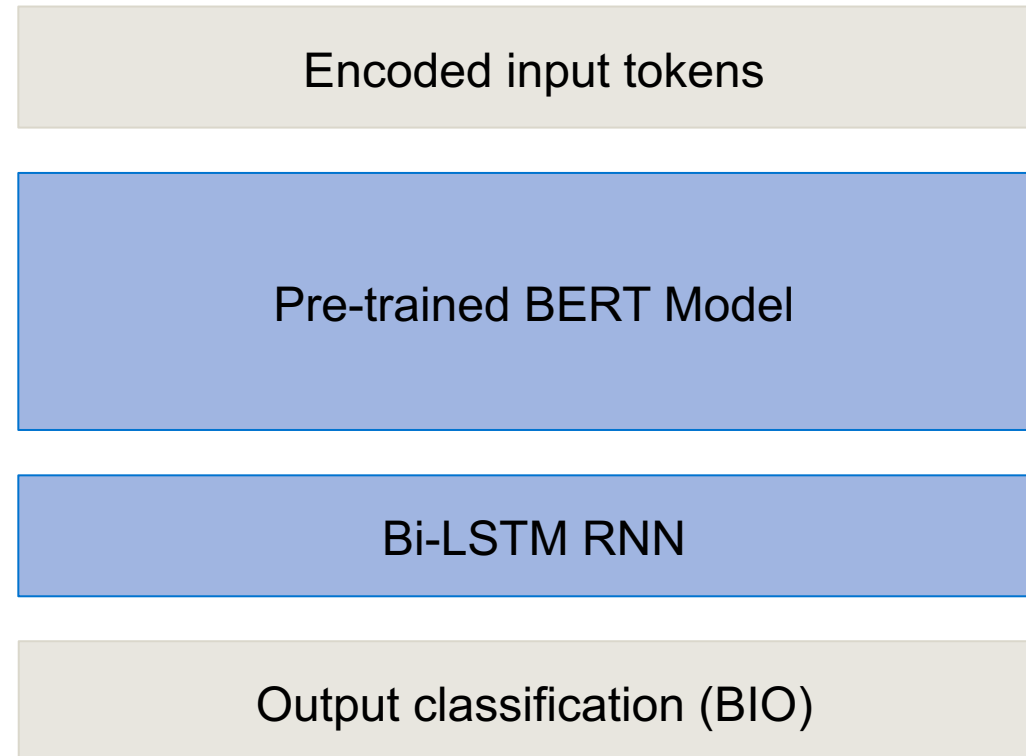
Natural party

Step 1: Dataset Preparation - Overview



Step 2: ML Model Training

Model Architecture



Performance of the model can be evaluated on

- General purpose named entity data sets (e.g. GermEval)
- Validation set from the labeled dataset
- Original / anonymized court decision pairs (e.g. from financial court)

Step 4: Building Anonymization Prototype

- Web app in browser
- Document upload
- Simple anonymization settings
- Model serving on server side

Browser

Anonymization Tool Webapp

Document Upload

Settings

☒ Judge names

☐ Birth dates

☒ Revenue

Show only year ▼

Motivation

Problem Statement

Approach

Research Questions

Schedule

1. Which special named entity types are most relevant for anonymization of court decisions?
2. How can anonymizations be marked in machine-detectable form?
3. What performance can be reached using state-of-the-art pretrained NLP models on specialized NER tasks for legal anonymization?

Motivation

Problem Statement

Approach

Research Questions

Schedule

Schedule

	March	April	May	June	July	August	September
Literature Research		Start – 15th of May					
Data Preprocessing		15th of April – 15th of July					
Model Training			15th of May – 4th of October				
Web App Implementation						August	
Writing			15th of May – End				
Review					16th of Sep. – End		



Tom Schamberger

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289. 17132
Fax +49.89.289.17136

matthes@in.tum.de
wwwmatthes.in.tum.de

