Anonymization of German Legal Texts

Bachelor Thesis - Kickoff

Tom Schamberger, Kickoff 18th of June 2019

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

Outline

Motivation

Problem Statement

Approach

Research Questions

Schedule

NLP for Legal Documents (today)

Manual Anonymization

NLP

Original legal texts

- Court decisions
- Contracts
- Accusatorial texts

➢ Not published

Anonymized legal texts

- No personal information

➢ Published

Knowledge

- Legal research
- Contract review
- Document automation
- Legal advice
- etc.

Outline

Motivation

Problem Statement

Approach

Research Questions

Schedule

Anonymization of Legal Texts

Example for court decisions:

Anonymization

6. Den Angeboten auf der Webseite ... liegt das Konzept der Produktdetailseite zugrunde. Dabei wird für jedes über die ...-Plattform angebotene Produkt jeweils nur eine Produkt-detailseite angezeigt; jedes Produkt enthält eine spezifische ...-Produktidentifikations-nummer (...) zugewiesen.

- **Expensive manual** anonymization process
  - ➢ Leads to rare publications of legal texts
  - ➢ Results in few data sets
- Development of an **automated anonymization process**
- Only anonymized data sets available
  - ➢ Anonymization training **without non-anonymized data sets**

Outline

Motivation

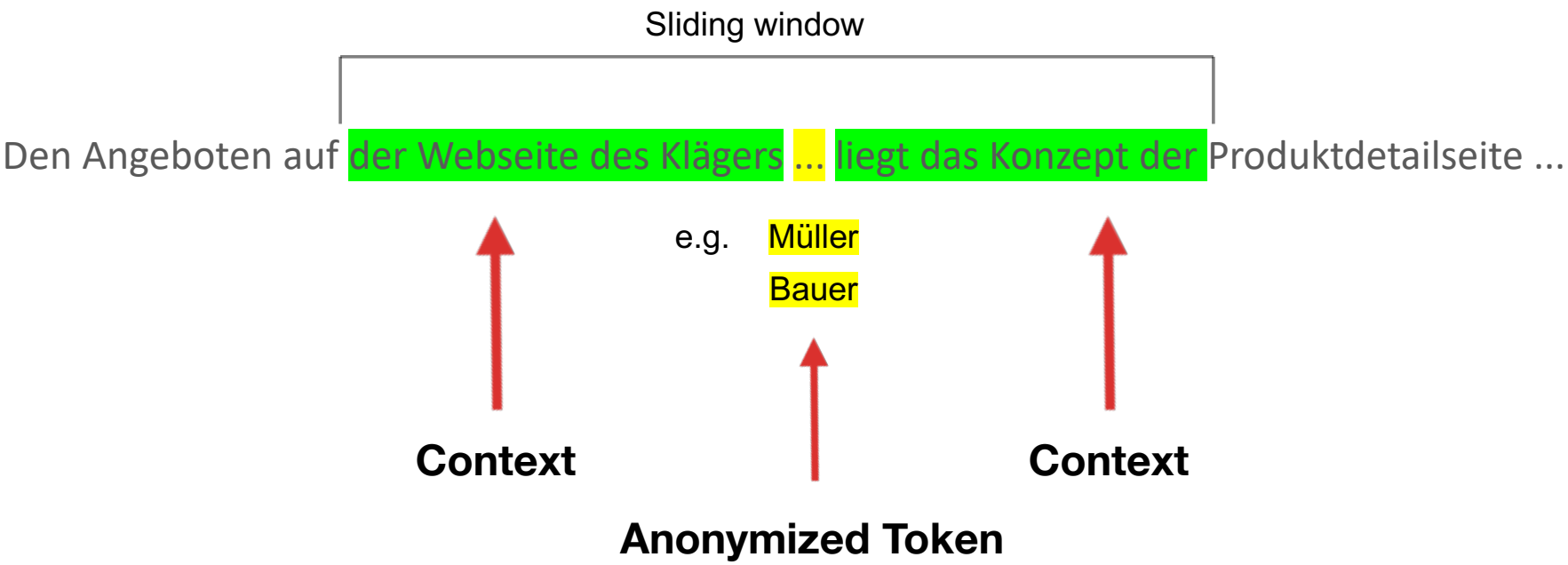
Problem Statement

Approach

Research Questions

Schedule

Context-based Analysis



- Statement: **Sensitivity** of token **depends only on context**, not the actual content itself
- Anonymized token in legal texts are annotated (e.g. by "...") and can be detected

➢ Anonymization model may be **trained using anonymized data**

Source: https://www.hrr-strafrecht.de/hrr/archiv/11-10/index.php?sz=6

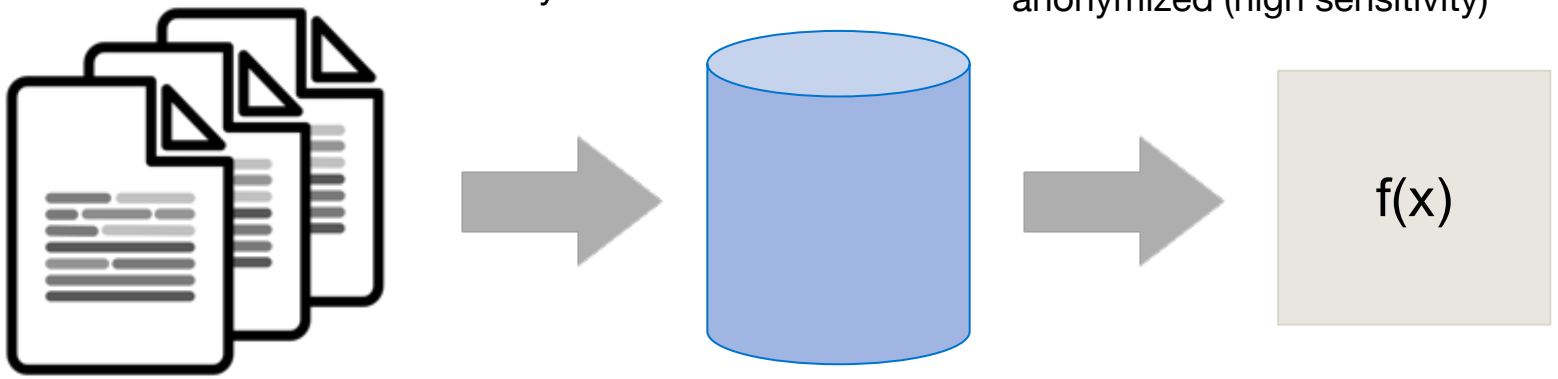
NLP: Sensitivity Prediction of Token

Preprocessing

- Text extraction
- Tokenization
- Anonymization Detection

Model Training

- Embedding of token context without the token itself
- Prediction if the token has been anonymized (high sensitivity)



Anonymized legal texts

- Raw data as PDF/RTF
- e.g. Court Decisions of ‘LG Munich’

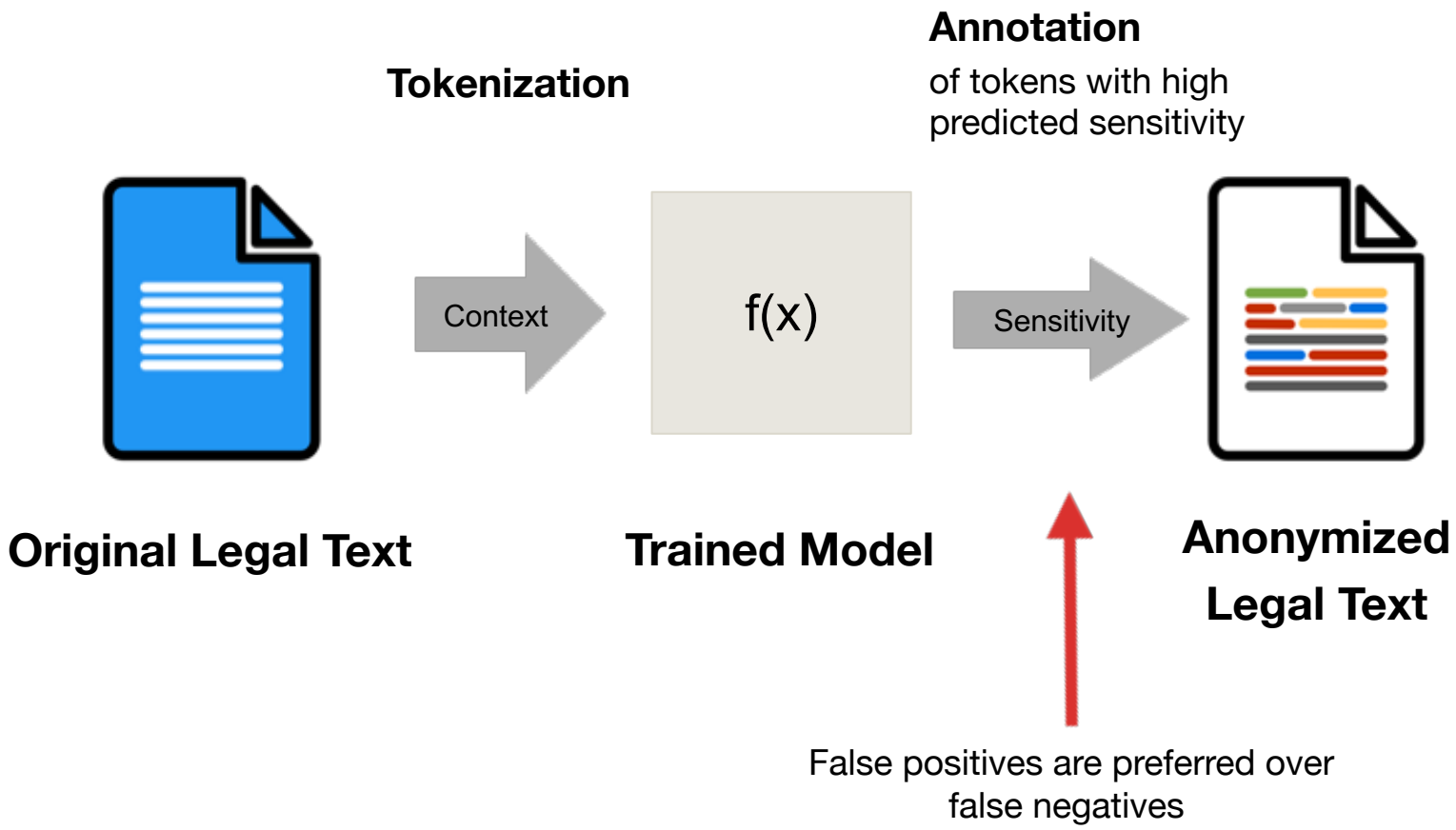
Tokenized Corpus

- Annotated anonymized tokens like ‘...’ or ‘B.’

Trained Model

- Inputs: Context
- Outputs: **Sensitivity**

Application of the Trained Model



Approach

1. **Literature research**
   - existing suitable NLP models
   - existing frameworks and implementations

2. **NLP model training** on anonymized court decisions
   - Fetching and preprocessing of court decision documents
   - Model training and hyperparameter fitting

3. **Evaluation of the model**
   - using court decision examples
   - using rewritten anonymized legal texts (e.g. using randomized NE replacement)

4. **Application**
   - Implementation of the anonymization process using trained NLP model

Outline

Motivation

Problem Statement

Approach

Research Questions

Schedule

Research Questions

Is the textual context of a token within legal texts enough to predict the sensitivity of this token?

- Classical anonymization solutions in other domains additionally use NER (named entity recognition)
  - Improves stability
  - But there is no non-anonymized data

- Additional data, which may be taken into account:
  - the **commonness** of words in legal texts
    - e.g. words like *mit* or *Kläger* probably have a low sensitivity
  - **document metadata** such as title, court, etc.

Research Questions

How can placeholders be detected in anonymized legal documents?

- Examples for **possible anonymization patterns** for the word ‘Schamberger’:
  - …
  - s…
  - S...
  - “S”
  - “S...”
  - s
  - S

Research Questions

Which machine learning approaches be used to automate anonymization using only anonymized data?

- Examples for contextual embeddings: **BERT**, ELMO, GPU, CMU
- Alternative classical embedding: Word2vec, GloVe

- Possible architectures: Convolution, RNN, Attention

- Possible sliding window sizes:
  - **Fixed number of words**
  - Full sentences
  - Full paragraphs

Outline

Motivation

Problem Statement

Approach

Research Questions

Schedule

Schedule

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