Anonymization of German Legal Texts

Bachelor Thesis - Final

Tom Schamberger, Final 4th of November 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

Conclusion

Future Work

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

Conclusion

Future Work

Anonymization of Legal Texts

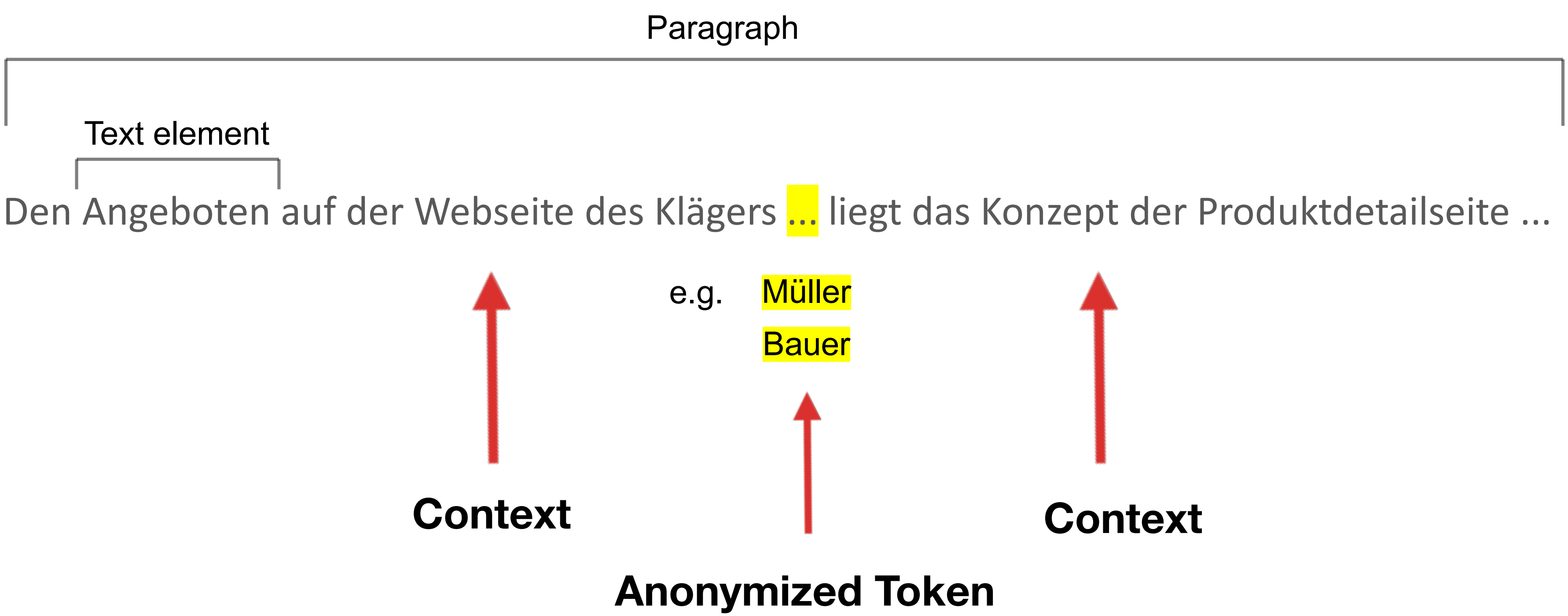
Example for court decisions:

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
- Use ML-based NLP to **automate sensitivity classification**
- Only anonymized data sets available
  - ➢ Anonymization training **without non-anonymized data sets**

Context-based Analysis



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

➢ Anonymization model is **trained using anonymized data**

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

Outline

Motivation

Problem Statement

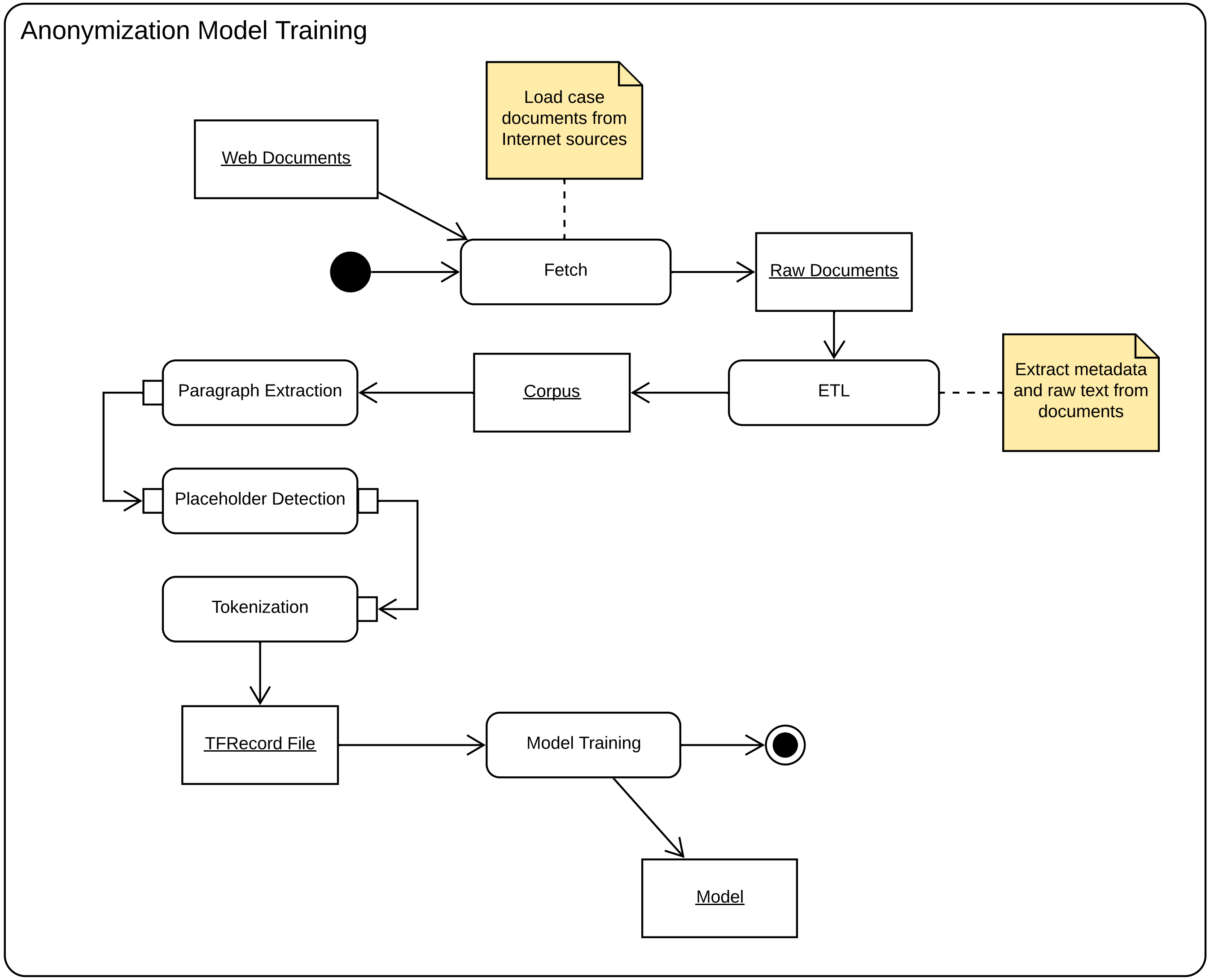
Approach

Research Questions

Conclusion

Future Work

NLP: Model Training



Text Corpus

- 1400 German anonymized court decisions of the state court in Munich (LG Munich)
- **Training corpus**
  - Document count: 1,220
  - Text element count: 4,181,266
  - Anonymized element count: 33,779
- **Test corpus** (rewritten anonymized documents)
  - Document count: 13
  - Text element count: 40,612
  - Anonymized element count: 358 (Unique: 123)

Text Corpus: Assumptions

- **Assumptions**
  - All anonymized references have been neutralized using placeholders without further modification
  - Placeholders are easily distinguishable from other text fractions
  - References have been replaced in such a way that the meaning of the text is retained
  - All documents have been anonymized using consistent and legally defined rules

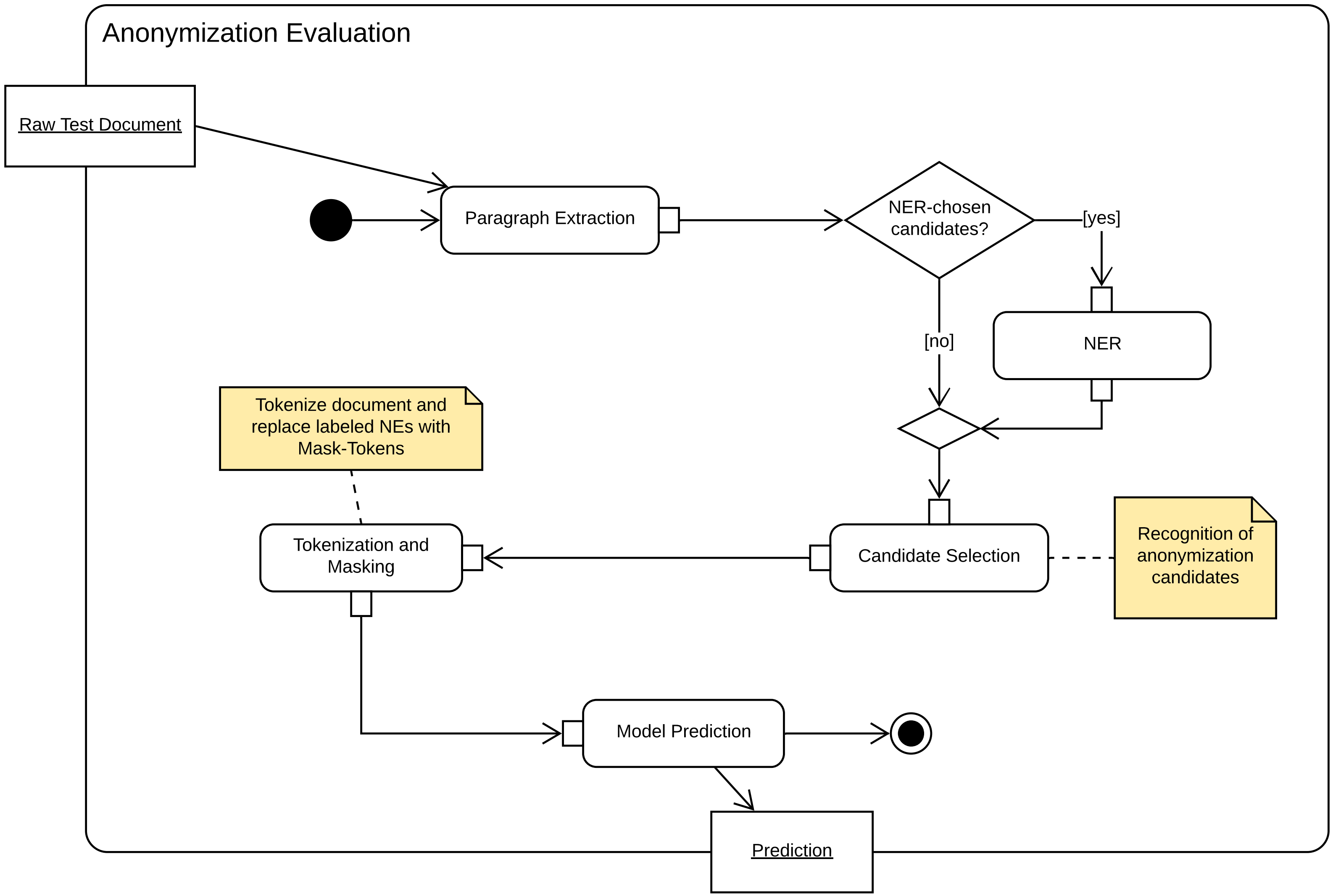
Model Evaluation

1. **Without candidates**
   - The model is applied on the whole input sequence without pre-selection

2. **Using random candidates**
   - Randomly chosen text fractions are masked
   - Aims to evaluate the ability to distinguish random text fractions w.r.t. sensitivity

3. **Using NER-detected candidates**
   - Named entities (NE) are detected and masked
   - Aims to evaluate the ability to distinguish NEs w.r.t. sensitivity

NLP: Sensitivity Prediction of Token



Outline

Motivation

Problem Statement

Approach

Research Questions

Conclusion

Future Work

Research Questions

How can placeholders be detected in anonymized legal documents?

- **Rule-based approach**
  - Distinction between “obvious” and “potential” placeholders
  - “Obvious” placeholders are specially marked e.g. using cites (e.g. *“a.”*)
  - "Potential” placeholders may alternatively stand for:
    - Omissions within cites such as testimonies
    - Abbreviations (e.g. *i.d.R.*)
    - References to pages or appendices (e.g. *siehe Anhang A 34*)
    - References to laws (e.g. *§ 8 Abs. 3 Ziffer 2 UWG*)
  - Detection using sliding window (size 3) and Regex patterns

Research Questions

How can placeholders be detected in anonymized legal documents?

- Evaluation using manual placeholder replacement in test corpus
- **Evaluation results**:

Accuracy	99.9
Precision	95.9
Recall (Sensitivity)	98.8
Specificity	99.9
F1 Score	97.0

Research Questions

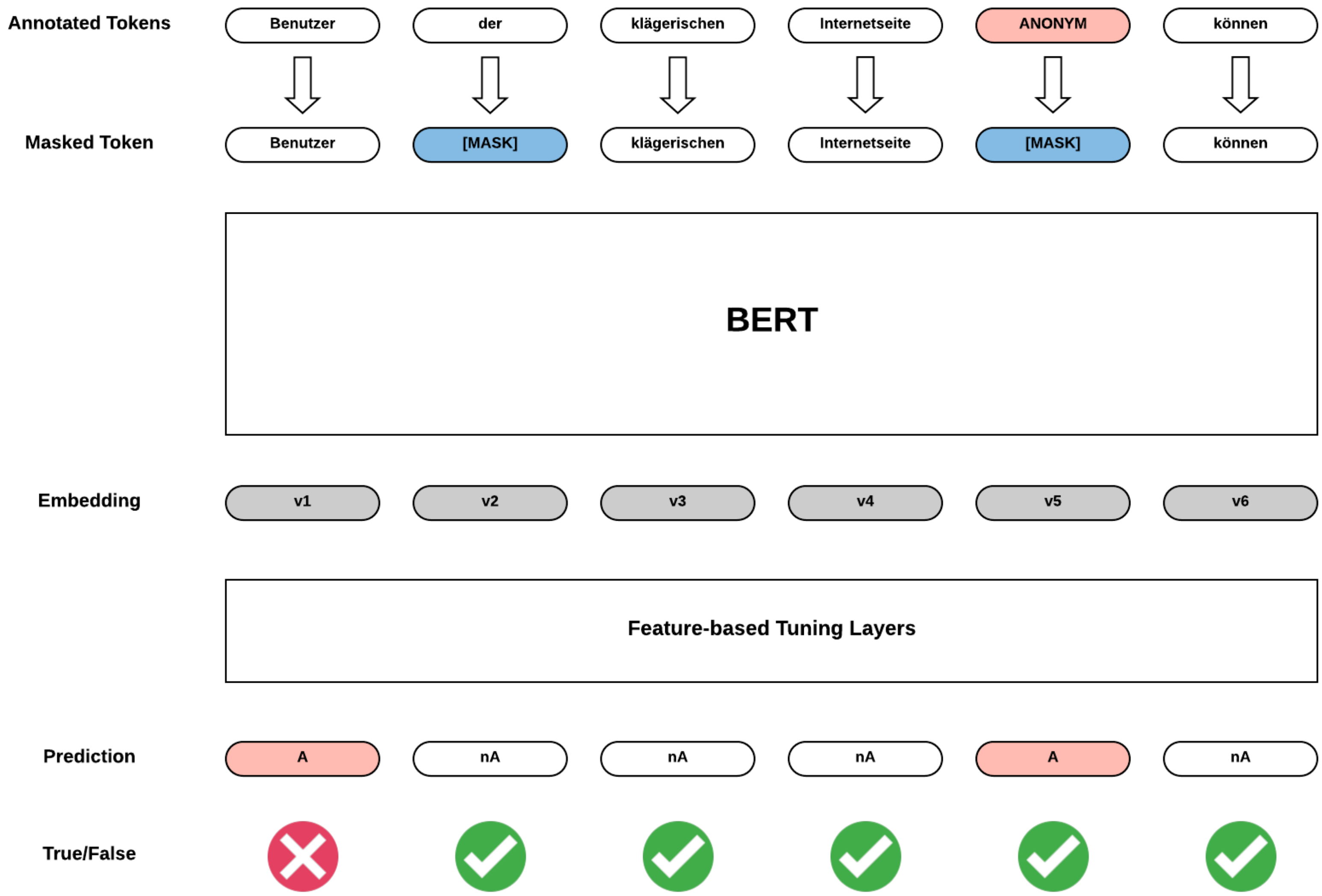
Which machine learning approaches are suitable to automate anonymization using only anonymized data?

- ➢ Evaluation of multiple deep learning approaches

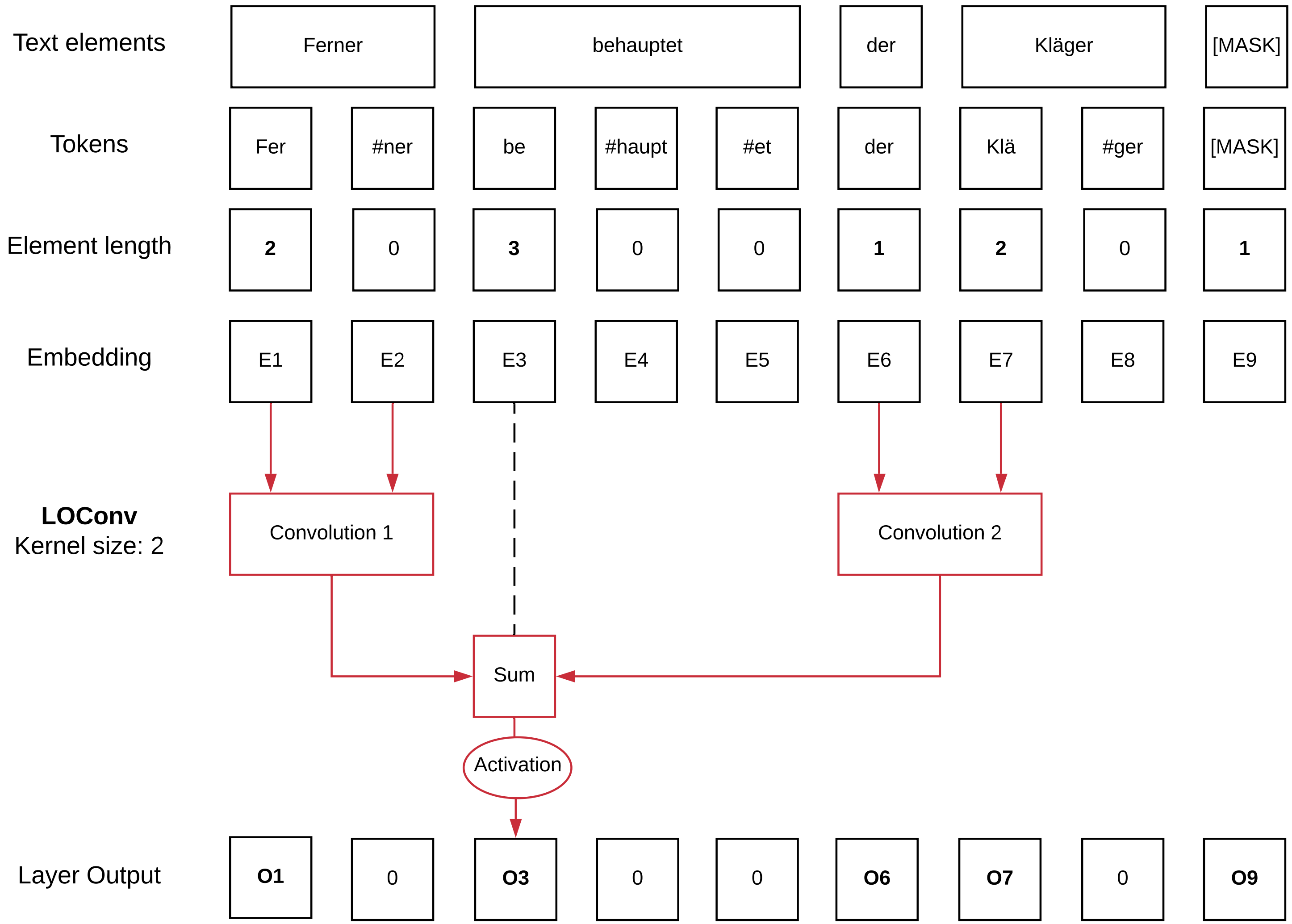
- **Embeddings**
  - BERT as a masked language model (masked LM)
  - GloVe word embeddings
    - ➢ In this work: Combination of universal pre-trained and specifically trained GloVe embeddings

- **Architectures**
  - Convolutional Layers
  - BiLSTM RNNs
  - BERT fine-tuning
  - Custom architectures for contextual learning:
    - LOConv
    - LOBiRNN

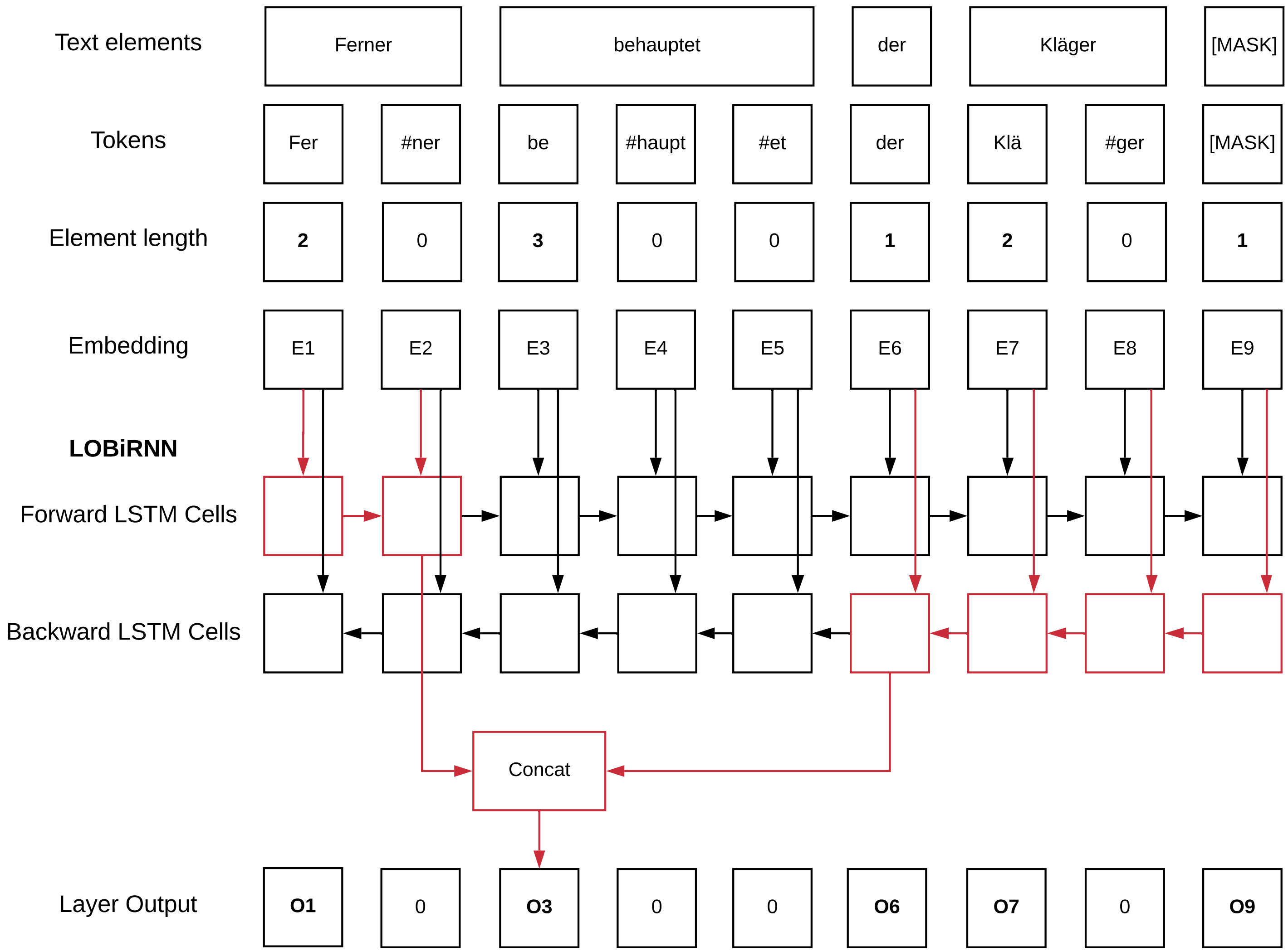
Architectures: BERT as a Masked LM



Architectures: Leave-Out Convolution (LoConv)



Architectures: Leave-Out RNN (LOBiRNN)



Architectures: Overview

- Evaluated Architectures*:

Variant-Ref.	Embedding	Type	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
FT	BERT	Fine-Tune**	-	-	-	-	-
RNN1	BERT	RNN	BiLSTM 128	BiLSTM 128	-	-	-
RNN2	BERT	RNN	BiLSTM 256	BiLSTM 256	1x128	1x64	-
RNN3	BERT	RNN	BiLSTM 128	BiLSTM 128	BiLSTM 128	1x128	1x64
RNN4	BERT	RNN	BiLSTM 512	BiLSTM 512	1x128	1x64	-
LOConv1	BERT	LOConv	32x256 (lo)	1x256	1x128	1x64	-
Dense	BERT	Dense	1x256	1x256	-	-	-
LOConv2	GloVe	LOConv	1x64	32x256 (lo)	1x128	1x64	-
LOConv3	GloVe	LOConv	1x128	32x512 (lo)	1x256	1x128	1x64
LOConv4	GloVe	LOConv	1x64	64x256 (lo)	1x128	1x64	-
LORNN1	GloVe	LOBiRNN	256 LSTM	1x256	1x128	-	-
LORNN2	GloVe	LOBiRNN	512 LSTM	512 LSTM	1x512	1x256	-
LORNN3	GloVe	LOBiRNN	512 LSTM	512 LSTM	512 LSTM	1x512	1x256

* Selected for presentation (a complete list can be found in the thesis)
** Fine-tuned using the BERT model itself (instead of feature-based training)

Research Questions

Is the textual context of text elements enough information to predict sensitivity in German legal texts?

- Evaluation results without candidates

Embedding	Variant	Candidates*	Masking**	Epochs	Test Recall	Test Prec	Test Acc
BERT	FT	YES	0.0/1.0/0.0	4	58.1	**68.9**	**99.352**
BERT	RNN1	NO	0.1/0.9/0.0	10	24.9	59.7	99.126
BERT	RNN1	YES	0.0/1.0/0.0	10	88.4	15.9	95.459
BERT	RNN2	YES	0.0/1.0/0.0	10	**90.0**	25.0	97.343
BERT	RNN3	YES	0.0/1.0/0.0	15	85.4	27.6	97.735
BERT	LOConv1	NO	0.0/1.0/0.0	8	27.0	54.1	99.088
GloVe	LOConv2	NO	1.0/0.0/0.0	50	35.9	63.6	99.196
GloVe	LOConv3	NO	1.0/0.0/0.0	50	41.9	64.9	99.232
GloVe	LOConv4	NO	1.0/0.0/0.0	50	33.0	65.6	99.198
GloVe	LORNN1	NO	1.0/0.0/0.0	50	14.3	47.3	99.034
GloVe	LORNN2	NO	1.0/0.0/0.0	50	15.1	63.6	99.111
GloVe	LORNN3	NO	1.0/0.0/0.0	50	14.6	66.7	99.119

* Refers to the use of candidates during training
** Format: Real/Random/Masked

Validation-Test Gap

- Masking of the training input leads to a performance decrease during evaluation on the test data set (not masked)
- ➢ Solution:
  - Candidate selection, randomized masking for BERT embeddings
  - LO-Architectures for Glove word embeddings

Embedding	Variant	Candidates*	Masking**	Test Recall	Test Prec	Val Recall	Val Prec
BERT	FT	YES	0.0/1.0/0.0	58.1	68.9	90.4	94.9
BERT	RNN1	NO	0.1/0.9/0.0	24.9	59.7	99.4	37.0
BERT	RNN1	YES	0.0/1.0/0.0	88.4	15.9	87.9	70.4
BERT	RNN2	YES	0.0/1.0/0.0	90.0	25.0	89.3	80.3
BERT	RNN3	YES	0.0/1.0/0.0	85.4	27.6	89.0	82.0
BERT	LOConv1	NO	0.0/1.0/0.0	27.0	54.1	97.7	25.5
GloVe	LOConv2	NO	1.0/0.0/0.0	35.9	63.6	27.4	56.2
GloVe	LOConv3	NO	1.0/0.0/0.0	41.9	64.9	33.4	54.0
GloVe	LOConv4	NO	1.0/0.0/0.0	33.0	65.6	28.0	56.0
GloVe	LORNN1	NO	1.0/0.0/0.0	14.3	47.3	15.8	40.1
GloVe	LORNN2	NO	1.0/0.0/0.0	15.1	63.6	18.0	52.0
GloVe	LORNN3	NO	1.0/0.0/0.0	14.6	66.7	17.1	52.8

* Refers to the use of candidates during training
** Format: Real/Random/Masked

Evaluation with Random Candidates

- Candidates are randomly chosen, but positives are always included
- Canidate labeling probability w.r.t. text elements: 2.5%

➢ Increases positive/negative ratio from 1% to 28%

- Evaluation results:

Embedding	Variant	Candidates*	Masking**	Epochs	Test Recall	Test Prec	Test Acc
BERT	FT	YES	0.0/0.0/1.0	2	97.6	92.6	97.084
BERT	RNN2	YES	0.0/0.0/1.0	10	93.8	88.1	94.776
BERT	RNN3	YES	0.0/0.0/1.0	10	91.6	81.7	92.155
BERT	RNN4	YES	0.0/0.0/1.0	15	93.0	86.6	94.051

* Refers to the use of candidates during training
** Format: Real/Random/Masked

- Due to the decreased positive/negative ratio, the precision of the fine-tuned BERT model, 92.6 %, results in a total precision of approximately 23.6%*.

* can be shown using simple probability calculus, see A.2 in thesis

Downside of Pure Contextual Analysis

- In many cases, contextual analysis of text elements often determines correctly the sensitivity of objects
- But those objects are often mentioned as internal references to named entities
- Example*:

Unstreitig hat ein Mitarbeiter der <<Beklagten>> dem Kunden <<Fausner>> eine Krankenversicherung angeboten. Dass die angerufene Telefonnummer für die Firma des <<Kunden>> in einem Branchenverzeichnis eingetragen wäre, behauptet die <<Beklagte>> nicht, sie mutmaßt dies nur.

➢ Using only context, sensitive references expressed as NEs and internal references to sensitive NEs are often hardly distinguishable

* <<>>-Notation refers to positive labeling

Evaluation with NER-chosen Candidates

- Candidates are chosen using a NER model
- ➢ Results:

Embedding	Variant	Candidates*	Masking**	Epochs	Test Recall	Test Prec	Test Acc
BERT	FT	YES	0.0/0.0/1.0	2	77.3	57	68.979
BERT	RNN2	YES	0.0/0.0/1.0	10	79.1	68.9	76.941
BERT	RNN3	YES	0.0/0.0/1.0	10	72.4	68.2	74.067
BERT	RNN4	YES	0.0/0.0/1.0	15	74.7	66.4	75.788

* Refers to the use of candidates during training
** Format: Real/Random/Masked

- NER model recognized only 80.5% of positive named entities representing an upper limit for recall
- Some domain-specific NE-types are not supported

Outline

Conclusion

- The sensitivity of text elements depends on both the context and the fact that the element is a names entity
- A specialized NER model is required to further improve anonymization results
- Placeholders in legal documents can be detected using rule-based algorithms, but more anonymization consistency is desirable
- Both quality and quantity of the legal documents must be further improved to overcome the need for a pre-trained language models

Outline

Future Work

- Improvements on quality and quantity of the legal training corpus
  - Reduce anonymization inconsistencies and errors
  - Collect and verify more legal documents from other courts and law firms

- Specialized NER
  - Support for more named entity types such as brands, account numbers, etc.
  - The type of NEs does not need to be detected
  - High recall is desirable

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