Efficient and Accurate Annotation of Large Text Corpora Using Representative Class Archetypes

Markus Löhde, B.Sc. Informatics

04.03.2023, Kick-Off

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

Introduction to CreateData4AI

Extrapolation

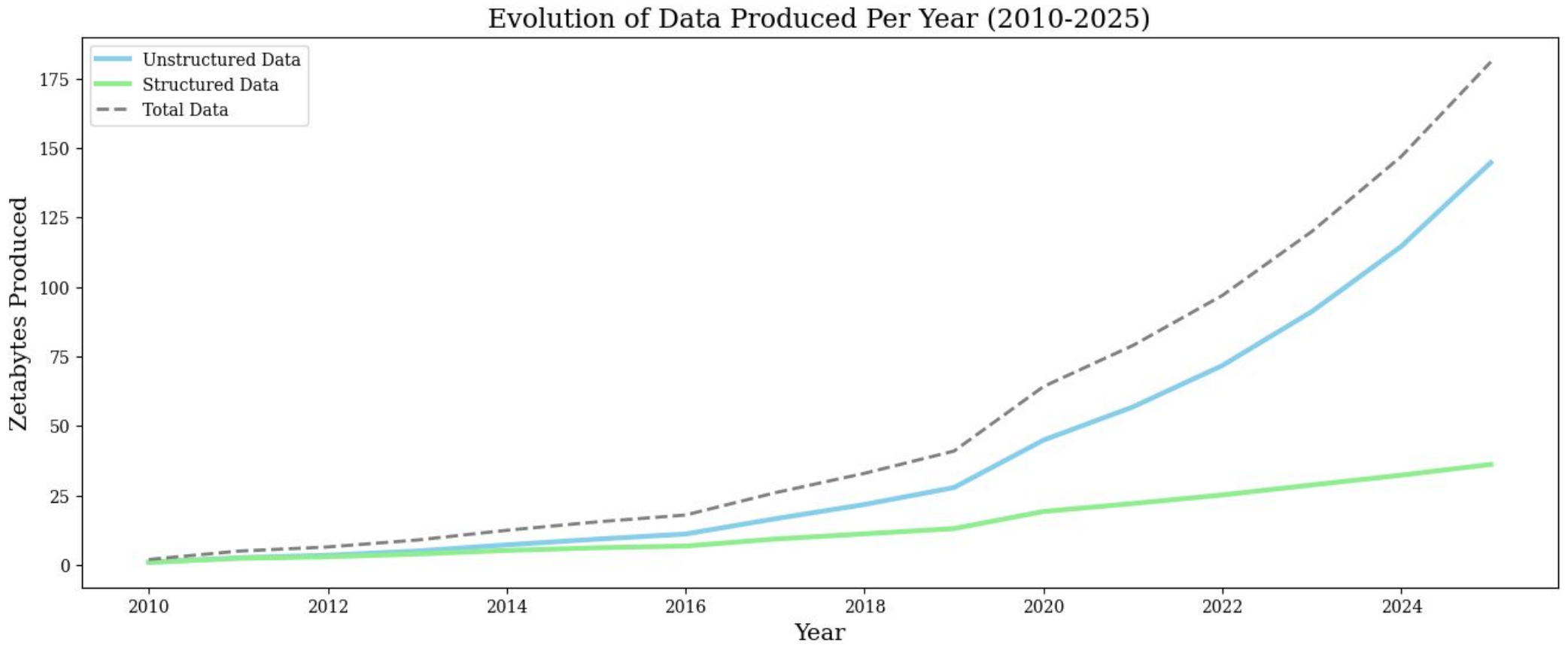
Initial Findings

Proposed Research Questions

Tasks & Timeline

Evolution of Data Creation

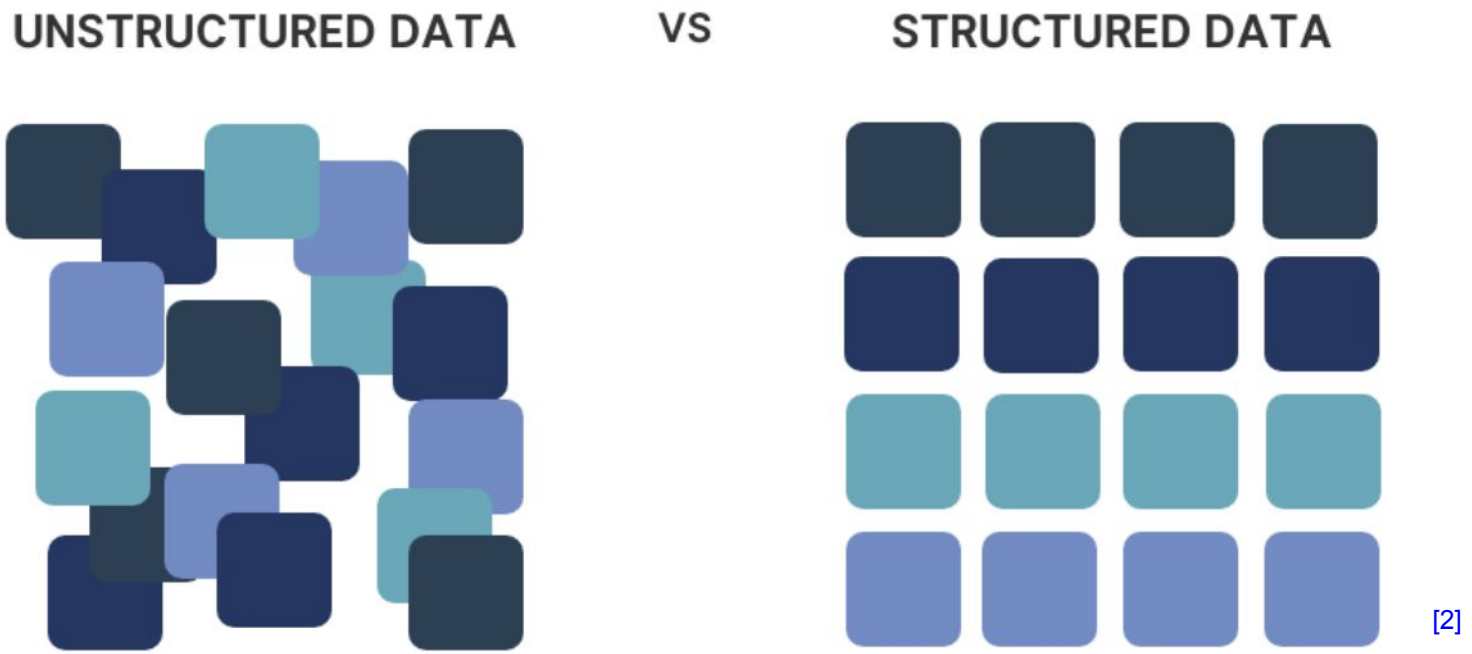
- With data creation increasing exponentially, we expect to produce 150 zetabytes globally in 2024.
- However ~80% of that data will be unstructured!



[1]

The Value of Structured Data

- Structuring unstructured data is still human-dependent and resource-intense
- Automating that process will allow especially smaller organizations to...
  - extract valuable insights from their data
  - train new models
  - enhance current model performance



[2]

CreateData4AI (CD4AI)

- In the CD4AI project we aim to develop a data-annotation pipeline with a human in the loop
- Our pilot project deals with a 3 million row dataset from the german trade register that details the purpose of companies
- The companies need to be categorized into 21 classes, corresponding to the 21 economic sectors defined by the german ministry of statistics

Input Data

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.
Rental Bau GmbH	Durchführung der Sanierung, Verkauf, Verwaltung von Immobilien oder grundstücksgleichen Rechten.
...	...

class	class_description
A	Land- und Forstwirtschaft, Fischerei
B	Bergbau und Gewinnung von Steinen und Erden
...	...

CD4AI Pipeline

Keyword Extraction → Context Windows → Extrapo-lation

Output

CreateData4AI (CD4AI)

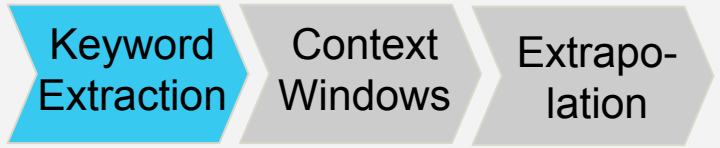
- In the CD4AI project we aim to develop a data-annotation pipeline with a human in the loop
- Our pilot project deals with a 3 million row dataset from the german trade register that details the purpose of companies
- The companies need to be categorized into 21 classes, corresponding to the 21 economic sectors defined by the german ministry of statistics

Input Data

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.
Rental Bau GmbH	Durchführung der Sanierung, Verkauf, Verwaltung von Immobilien oder grundstücksgleichen Rechten.
...	...

class	class_description
A	Land- und Forstwirtschaft, Fischerei
B	Bergbau und Gewinnung von Steinen und Erden
...	...

CD4AI Pipeline



Output

CreateData4AI (CD4AI)

- In the CD4AI project we aim to develop a data-annotation pipeline with a human in the loop
- Our pilot project deals with a 3 million row dataset from the german trade register that details the purpose of companies
- The companies need to be categorized into 21 classes, corresponding to the 21 economic sectors defined by the german ministry of statistics

Input Data

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.
Rental Bau GmbH	Durchführung der Sanierung, Verkauf, Verwaltung von Immobilien oder grundstücksgleichen Rechten.
...	...

class	class_description
A	Land- und Forstwirtschaft, Fischerei
B	Bergbau und Gewinnung von Steinen und Erden
...	...

CD4AI Pipeline

Keyword Extraction → Context Windows → Extrapo-lation

"Bauernhof"

Output

CreateData4AI (CD4AI)

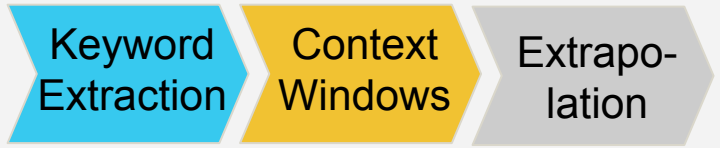
- In the CD4AI project we aim to develop a data-annotation pipeline with a human in the loop
- Our pilot project deals with a 3 million row dataset from the german trade register that details the purpose of companies
- The companies need to be categorized into 21 classes, corresponding to the 21 economic sectors defined by the german ministry of statistics

Input Data

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.
Rental Bau GmbH	Durchführung der Sanierung, Verkauf, Verwaltung von Immobilien oder grundstücksgleichen Rechten.
...	...

class	class_description
A	Land- und Forstwirtschaft, Fischerei
B	Bergbau und Gewinnung von Steinen und Erden
...	...

CD4AI Pipeline



Output

CreateData4AI (CD4AI)

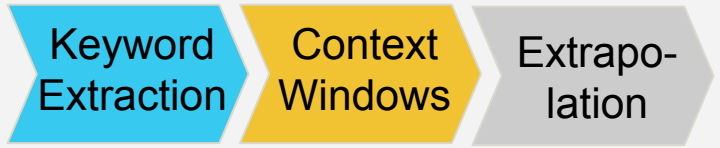
- In the CD4AI project we aim to develop a data-annotation pipeline with a human in the loop
- Our pilot project deals with a 3 million row dataset from the german trade register that details the purpose of companies
- The companies need to be categorized into 21 classes, corresponding to the 21 economic sectors defined by the german ministry of statistics

Input Data

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.
Rental Bau GmbH	Durchführung der Sanierung, Verkauf, Verwaltung von Immobilien oder grundstücksgleichen Rechten.
...	...

class	class_description
A	Land- und Forstwirtschaft, Fischerei
B	Bergbau und Gewinnung von Steinen und Erden
...	...

CD4AI Pipeline



"Betreiben eines Bauernhofs"

Output

CreateData4AI (CD4AI)

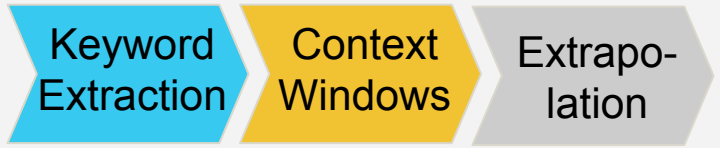
- In the CD4AI project we aim to develop a data-annotation pipeline with a human in the loop
- Our pilot project deals with a 3 million row dataset from the german trade register that details the purpose of companies
- The companies need to be categorized into 21 classes, corresponding to the 21 economic sectors defined by the german ministry of statistics

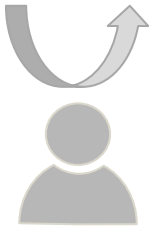
Input Data

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.
Rental Bau GmbH	Durchführung der Sanierung, Verkauf, Verwaltung von Immobilien oder grundstücksgleichen Rechten.
...	...

class	class_description
A	Land- und Forstwirtschaft, Fischerei
B	Bergbau und Gewinnung von Steinen und Erden
...	...

CD4AI Pipeline





Domain Expert

Output

CreateData4AI (CD4AI)

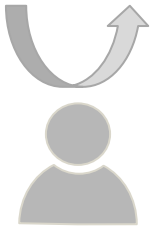
- In the CD4AI project we aim to develop a data-annotation pipeline with a human in the loop
- Our pilot project deals with a 3 million row dataset from the german trade register that details the purpose of companies
- The companies need to be categorized into 21 classes, corresponding to the 21 economic sectors defined by the german ministry of statistics

Input Data

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.
Rental Bau GmbH	Durchführung der Sanierung, Verkauf, Verwaltung von Immobilien oder grundstücksgleichen Rechten.
...	...

class	class_description
A	Land- und Forstwirtschaft, Fischerei
B	Bergbau und Gewinnung von Steinen und Erden
...	...

CD4AI Pipeline



Domain Expert

Output

CreateData4AI (CD4AI)

- In the CD4AI project we aim to develop a data-annotation pipeline with a human in the loop
- Our pilot project deals with a 3 million row dataset from the german trade register that details the purpose of companies
- The companies need to be categorized into 21 classes, corresponding to the 21 economic sectors defined by the german ministry of statistics

Input Data

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.
Rental Bau GmbH	Durchführung der Sanierung, Verkauf, Verwaltung von Immobilien oder grundstücksgleichen Rechten.
...	...

class	class_description
A	Land- und Forstwirtschaft, Fischerei
B	Bergbau und Gewinnung von Steinen und Erden
...	...

CD4AI Pipeline

Output

CreateData4AI (CD4AI)

- In the CD4AI project we aim to develop a data-annotation pipeline with a human in the loop
- Our pilot project deals with a 3 million row dataset from the german trade register that details the purpose of companies
- The companies need to be categorized into 21 classes, corresponding to the 21 economic sectors defined by the german ministry of statistics

Input Data

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.
Rental Bau GmbH	Durchführung der Sanierung, Verkauf, Verwaltung von Immobilien oder grundstücksgleichen Rechten.
...	...

class	class_description
A	Land- und Forstwirtschaft, Fischerei
B	Bergbau und Gewinnung von Steinen und Erden
...	...

CD4AI Pipeline

Output

legal_name	purpose	class
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.	H
Rental Bau GmbH	Durchführung der Sanierung, Verkauf, Verwaltung von Immobilien oder grundstücksgleichen Rechten.	L
...	...	...

CreateData4AI (CD4AI)

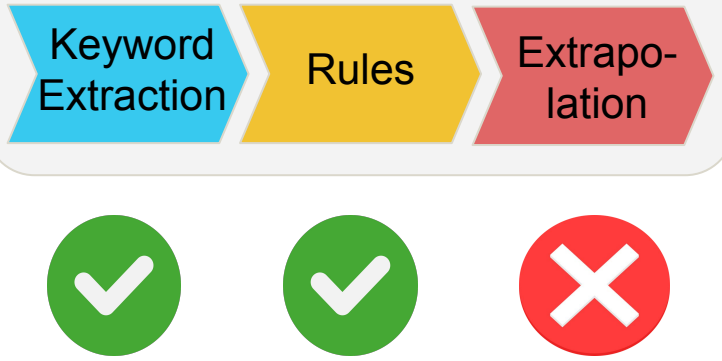
- In the CD4AI project we aim to develop a data-annotation pipeline with a human in the loop
- Our pilot project deals with a 3 million row dataset from the german trade register that details the purpose of companies
- The companies need to be categorized into 21 classes, corresponding to the 21 economic sectors defined by the german ministry of statistics

Input Data

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.
Rental Bau GmbH	Durchführung der Sanierung, Verkauf, Verwaltung von Immobilien oder grundstücksgleichen Rechten.
...	...

class	class_description
A	Land- und Forstwirtschaft, Fischerei
B	Bergbau und Gewinnung von Steinen und Erden
...	...

CD4AI Pipeline



Output

legal_name	purpose	class
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.	H
Rental Bau GmbH	Durchführung der Sanierung, Verkauf, Verwaltung von Immobilien oder grundstücksgleichen Rechten.	L
...	...	...

Outline

Introduction to CreateData4AI

Proposed Research Questions

Extrapolation

Initial Findings

Tasks & Timeline

Proposed Research Questions

Main Research Question:

How can current state-of-the-art NLP techniques be used to annotate large, domain-specific text corpora?

Supporting Research Questions:

1. What is the most efficient and accurate approach for partitioning a set of text documents into categories defined by specific context rules?
2. What NLP strategies are best suited for identifying the most representative sentences in a text document?
3. What is the best method to evaluate a system that annotates large, domain-specific text corpora?

Outline

Introduction to CreateData4AI

Proposed Research Questions

Extrapolation

Initial Findings

Tasks & Timeline

Extrapolation - A two stage process

- The rules will capture both the semantic and the syntactic core of their class
- Since we are dealing with big data (~3 million rows) we have to find an accurate yet efficient solution
- The idea is to first leverage the syntactic similarities of the rules to narrow down the search space
- Then in a second step, we want to use more advanced semantic based techniques to assign the final label
- We will use the following as a running example:

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.

class	description	rules
H	Verkehr und Lagerung	['personenbeförderung', 'fuhrgeschäft', 'betrieb einer spedition']
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'

Extrapolation - Stage 1

- The goal of the first stage is to narrow down the search space by utilizing the syntactic similarity between rules and documents
- So we want to find the top k classes (k << 21), according to the number of matches found through exact or fuzzy string matching

k = 2

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.

class	description	rules	
H	Verkehr und Lagerung	['personenbeförderung', 'fuhrgeschäft', 'betrieb einer spedition']	0
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']	0
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'	0

Extrapolation - Stage 1

- The goal of the first stage is to narrow down the search space by utilizing the syntactic similarity between rules and documents
- So we want to find the top k classes (k << 21), according to the number of matches found through exact or fuzzy string matching

k = 2

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.

class	description	rules	
H	Verkehr und Lagerung	['personenbeförderung', 'fuhrgeschäft', 'betrieb einer spedition']	0
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']	0
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'	0

Extrapolation - Stage 1

- The goal of the first stage is to narrow down the search space by utilizing the syntactic similarity between rules and documents
- So we want to find the top k classes (k << 21), according to the number of matches found through exact or fuzzy string matching

k = 2

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.

class	description	rules	
H	Verkehr und Lagerung	['personenbeförderung', 'fuhrgeschäft', 'betrieb einer spedition']	0
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']	0
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'	0

Extrapolation - Stage 1

- The goal of the first stage is to narrow down the search space by utilizing the syntactic similarity between rules and documents
- So we want to find the top k classes (k << 21), according to the number of matches found through exact or fuzzy string matching

k = 2

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.

class	description	rules	
H	Verkehr und Lagerung	['personenbeförderung', 'fuhrgeschäft', 'betrieb einer spedition']	1
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']	0
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'	0

Extrapolation - Stage 1

- The goal of the first stage is to narrow down the search space by utilizing the syntactic similarity between rules and documents
- So we want to find the top k classes (k << 21), according to the number of matches found through exact or fuzzy string matching

k = 2

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.

class	description	rules	
H	Verkehr und Lagerung	['personenbeförderung', 'fuhrgeschäft', 'betrieb einer spedition']	1
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']	1
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'	0

Extrapolation - Stage 1

- The goal of the first stage is to narrow down the search space by utilizing the syntactic similarity between rules and documents
- So we want to find the top k classes (k << 21), according to the number of matches found through exact or fuzzy string matching

k = 2

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.

class	description	rules	
H	Verkehr und Lagerung	['personenbeförderung', 'fuhrgeschäft', 'betrieb einer spedition']	1
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']	1
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'	0

Extrapolation - Stage 1

- The goal of the first stage is to narrow down the search space by utilizing the syntactic similarity between rules and documents
- So we want to find the top k classes (k << 21), according to the number of matches found through exact or fuzzy string matching

k = 2

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.

class	description	rules	
H	Verkehr und Lagerung	['personenbeförderung', 'fuhrgeschäft', 'betrieb einer spedition']	1
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']	1
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'	0

Extrapolation - Stage 1

- The goal of the first stage is to narrow down the search space by utilizing the syntactic similarity between rules and documents
- So we want to find the top k classes (k << 21), according to the number of matches found through exact or fuzzy string matching

k = 2

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.

class	description	rules	
H	Verkehr und Lagerung	['personenbeförderung', 'fuhrgeschäft', 'betrieb einer spedition']	1
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']	1
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'	0

Extrapolation - Stage 1

- The goal of the first stage is to narrow down the search space by utilizing the syntactic similarity between rules and documents
- So we want to find the top k classes (k << 21), according to the number of matches found through exact or fuzzy string matching

k = 2

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.

class	description	rules	
H	Verkehr und Lagerung	['personenbeförderung', 'fuhrgeschäft', 'betrieb einer spedition']	1
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']	1
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'	0

Extrapolation - Stage 1

- The goal of the first stage is to narrow down the search space by utilizing the syntactic similarity between rules and documents
- So we want to find the top k classes (k << 21), according to the number of matches found through exact or fuzzy string matching

k = 2

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.

class	description	rules	
H	Verkehr und Lagerung	['personenbeförderung', 'fuhrgeschäft', 'betrieb einer spedition']	1
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']	1
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'	0

Extrapolation - Stage 2

- With a smaller search space we can use more resource-intensive methods which leverage the semantic similarity between rules and documents
- In a first step, we will use to boil down the document to the most meaningful parts using e.g textRank [3]
- In a second step, we will use state-of-the-art transformer models, e.g. S-BERT[4] or setfit [5], to find the set of rules that best match our document's meaning and assign the corresponding class.

Extrapolation - Stage 2

Stage 1

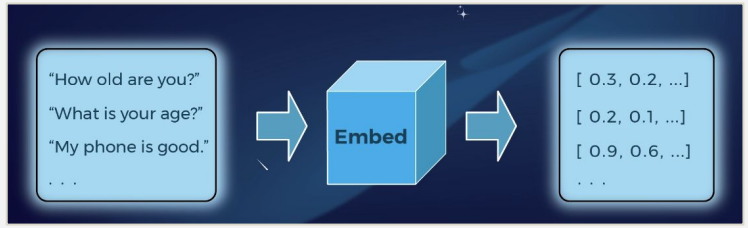
Sentence Ranking via e.g. **textRank**

legal_name	purpose
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.

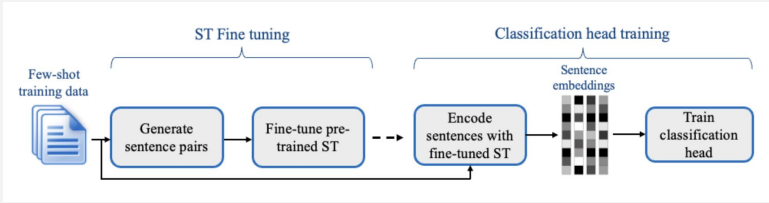
class	description	rules
H	Verkehr und Lagerung	['personenbeförderung', 'fuhrgeschäft', 'betrieb einer spedition']
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']

To be decided

Embeddings and Similarity Comparison



Fine-Tuned Sentences Transformer



[4]

legal_name	purpose	class
Wehle GmbH Spedition	Der Betrieb einer Spedition und eines Transportunternehmens.	H

Outline

Introduction to CreateData4AI

Proposed Research Questions

Extrapolation

Initial Findings

Tasks & Timeline

Initial Findings

- Disclaimer: We are currently still improving the process of creating the rules, so these results are very much preliminary.

Stage	Strategy	Computation Time per Document [2]	Accuracy [1]
1	Exact String Matching	~0.1 seconds	~67%
1	Fuzzy String Matching[3]	~0.5 seconds	~52%
2	S-BERT Embeddings	~0.2 seconds	~50%[4]

1. For Stage 1 we measure if the correct class is in the top 3 classes.
   For Stage 2 we measure if the final class is correct.
   Furthermore we used 33 documents with a balanced class distribution.
2. Measured on shared T4 GPU in Google Colab.
3. We used `partial_token_sort_ratio` from “thefuzz” [6]
4. Here we are limited by the accuracy of stage 1, so 0.5/0.67 = 74%

Outline

Introduction to CreateData4AI

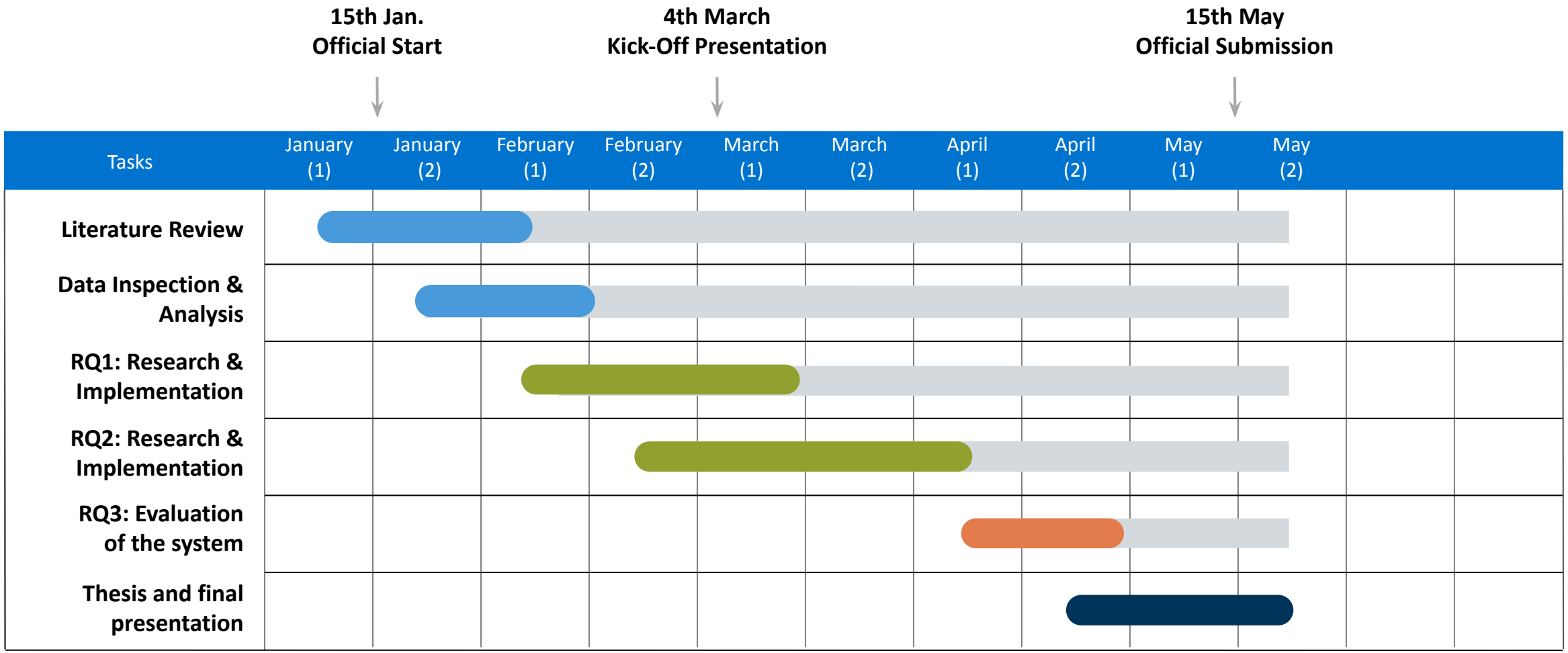
Proposed Research Questions

Extrapolation

Initial Findings

Tasks & Timeline

Tasks & Timeline



Prof. Dr.
Florian Matthes

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.tum.de

wwwmatthes.in.tum.de

References

[1] Data growth worldwide 2010-2025 | Statista

[2] Unstructured VS Structured Data: 4 Key Management Differences [Infographic] – DryvIQ

[3] Mihalcea, R., & Tarau, P. (n.d.). TextRank: Bringing Order into Texts. Department of Computer Science, University of North Texas. Retrieved from cs.unt.edu

[4] Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. arXiv:2004.09813v2 [cs.CL]. https://doi.org/10.48550/arXiv.2004.09813

[5] Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient Few-Shot Learning Without Prompts. arXiv. https://doi.org/10.48550/arXiv.2209.11055

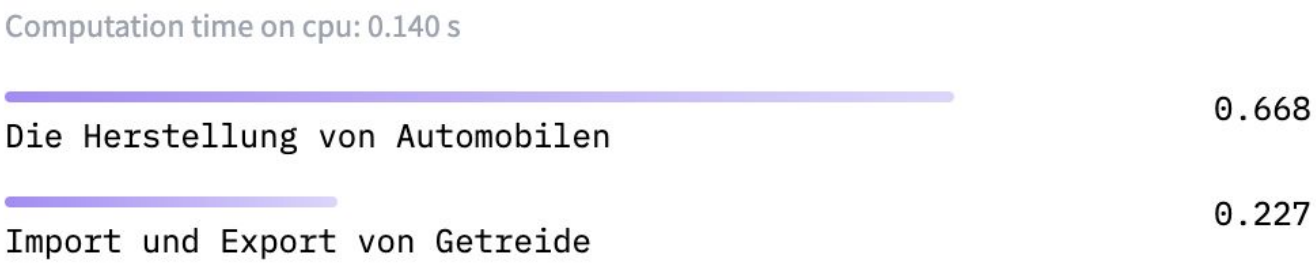
[6] https://github.com/seatgeek/thefuzz?tab=readme-ov-file

SetFit

- SetFit could be very useful for us because for our classification tasks sentences that are semantically similar in general, do not always belong to the same class.
- For example consider:

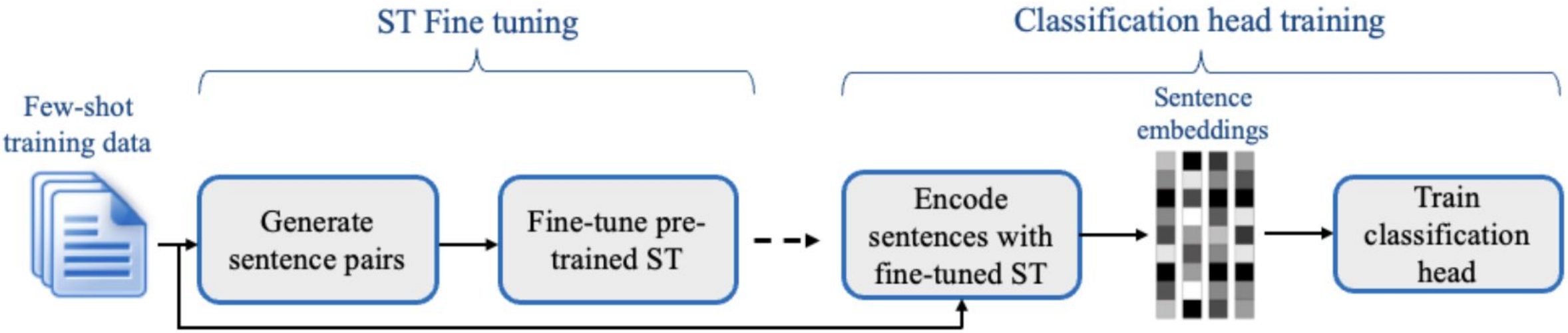
1. “Der Handel mit gebrauchten Automobilen.” → G: Handel
2. “Die Herstellung von Automobilen.” → C: Verarbeitendes Gewerbe
3. “Import und Export von Getreide” → G: Handel

- When comparing 1&2 and 1&3 this is the output of a sentence transformer not fine-tuned on our data:



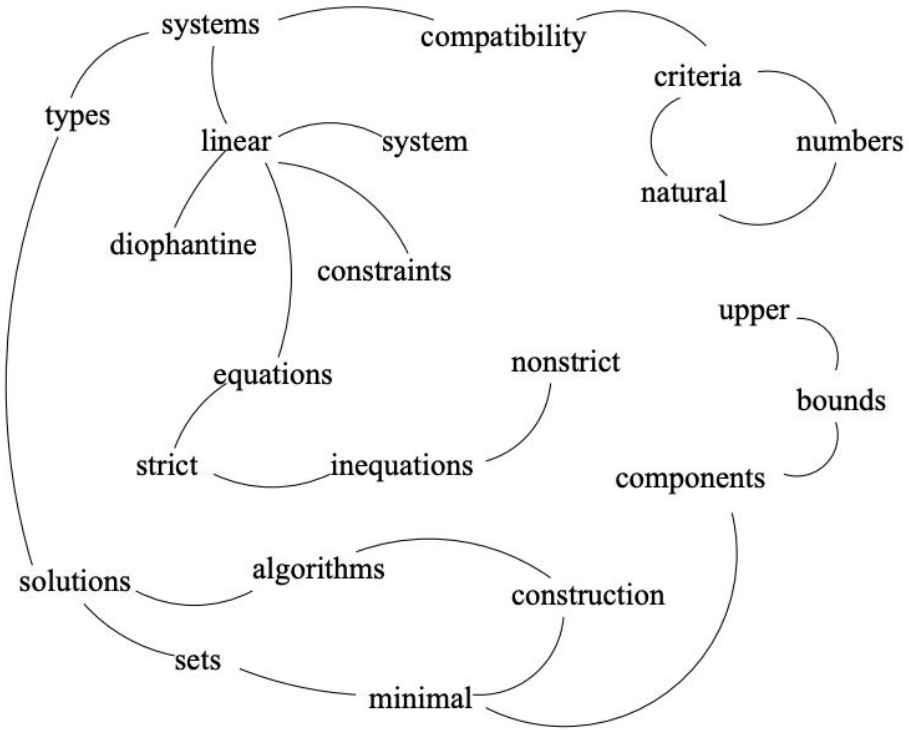
- SetFit allows us to efficiently fine-tune our sentence transformer, so that the produced embeddings are customized to our specific classification task.

SetFit

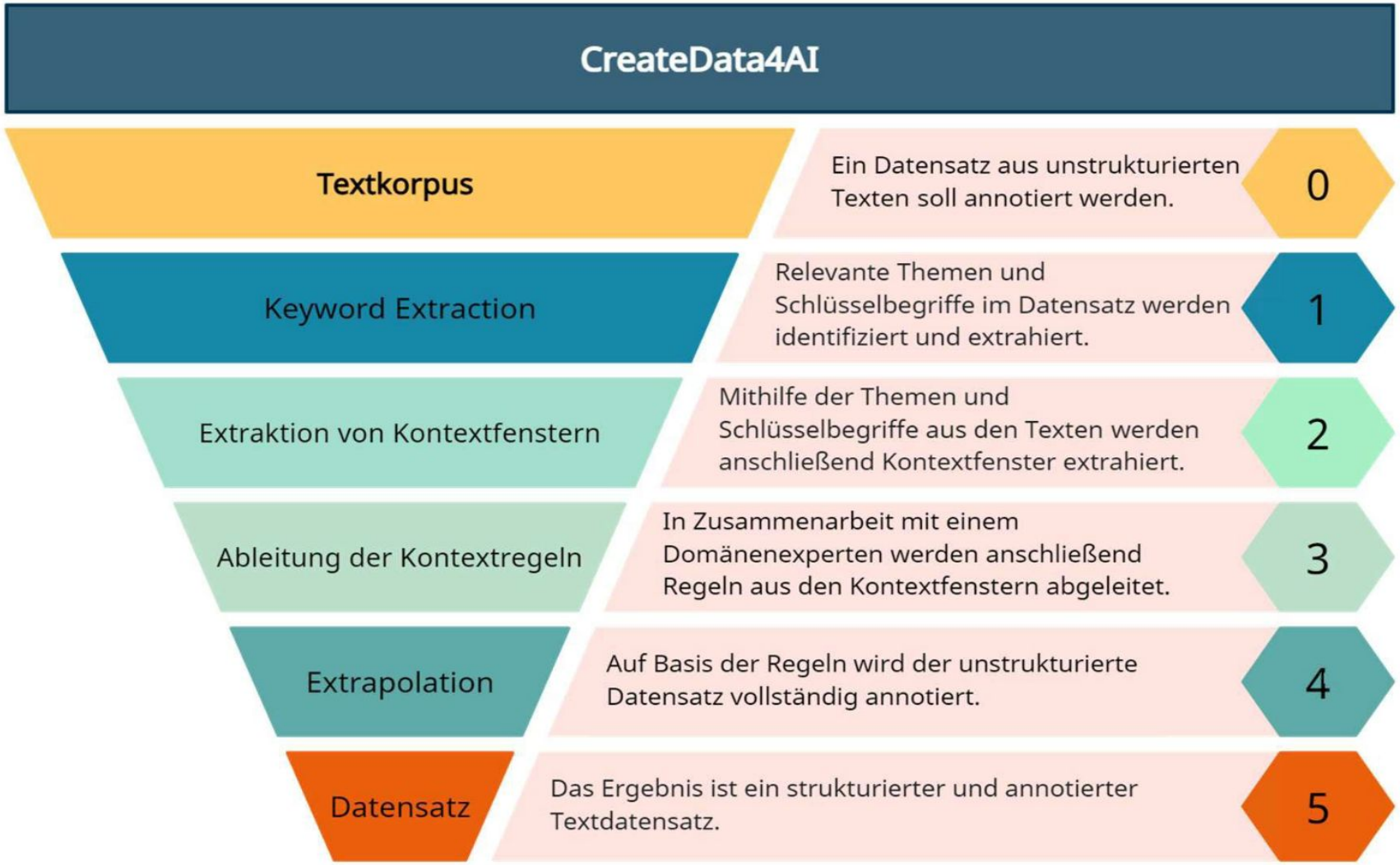

ST Fine tuning
Classification head training
Few-shot training data
Generate sentence pairs
Fine-tune pre-trained ST
Encode sentences with fine-tuned ST
Sentence embeddings
Train classification head

TextRank

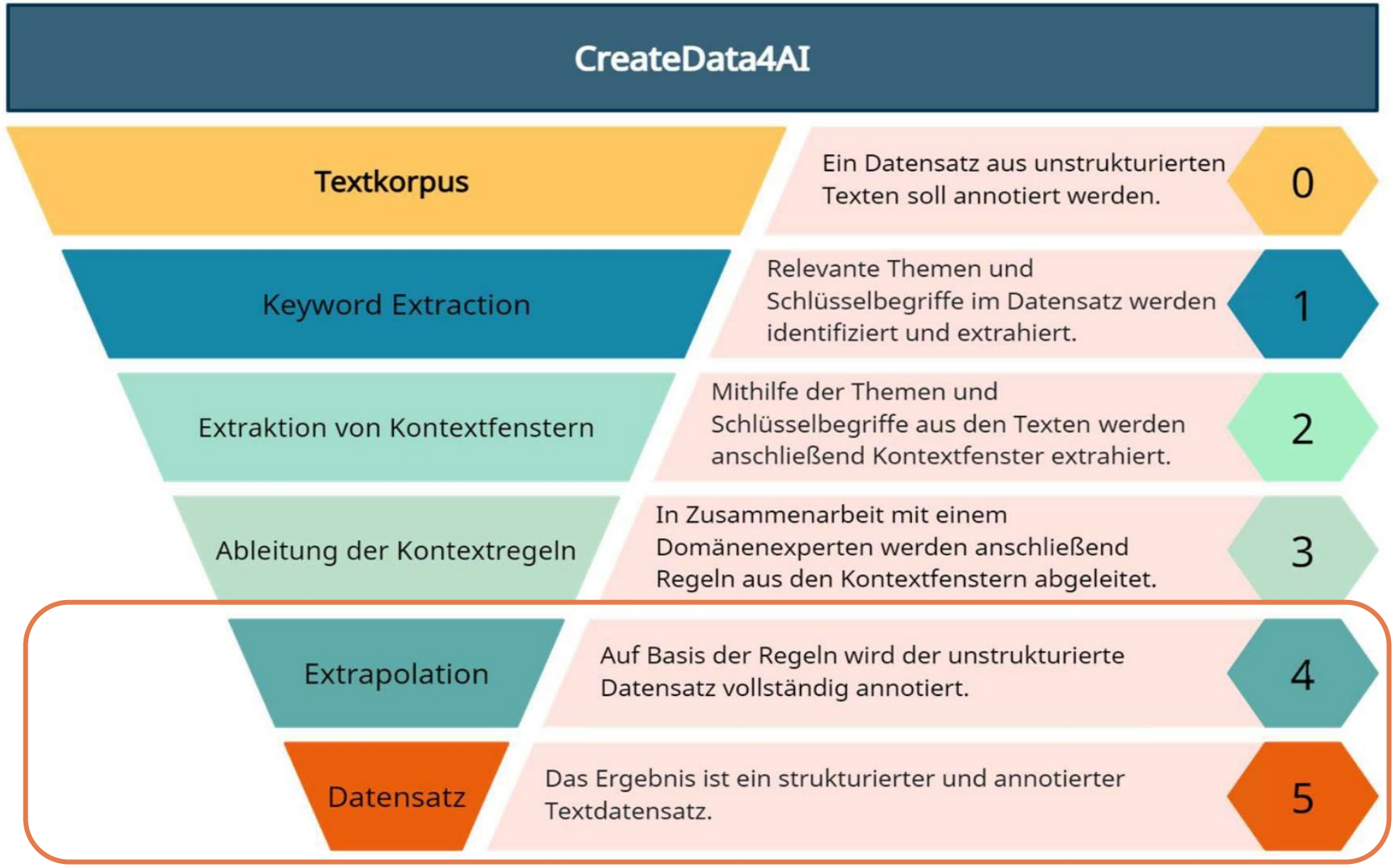
- textRank is a graph-based ranking model for sentences and keywords
- The graph is constructed according to either the content overlap of two sentences or the frequency of two words following each other in the document
- Once constructed, the pageRank algorithm is run to find the most important sentences/words



Overview of CreateData4AI



Overview of CreateData4AI



Process Flow

Input

Category rules and the unlabeled dataset

1 **Efficiently partitioning the dataset into categories using the rules**

2 **Finding the most relevant sentences in each document**

3 **Confidently labelling each document based on the rules**

Output

The original dataset with accurate labels

Efficiency ⟷ **Accuracy**

Technical details

- Efficient semantic embeddings
- Efficient, seeded clustering algorithms

Technical details

- Text Ranking Algorithms
- And some other stuff thats not so relevant

Technical details

- Highly accurate semantic embeddings
- Ensemble approach & Active Learning

Motivation for CreateData4AI (CD4AI)

328.77 million terabytes

The amount of data that is produced globally every day!

80% unlabeled

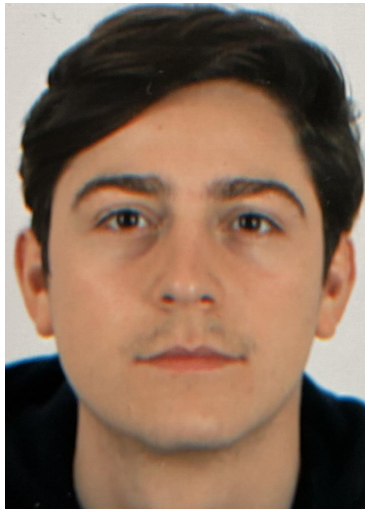
This leaves ~263 million terabytes of data unlabeled!

80/20 Rule

Highlights the fact that data scientists spend about 80% of their time preparing datasets.

Process Black Box

Personal Introduction



Curriculum Vitae

- Born in Hamburg, Germany
- Abitur in Hamburg at Gelehrtenschule des Johanneums (Grade: 1,4)
- B.Sc. in Computer Science at TUM
- Sport enthusiast, currently playing hockey at a professional level
- Interested in economics, productivity and AI

Relevant Experience

- **Preisenergie (Working Student) - 1.5 years:** Fundamentals of software engineering and architecture
- **Celonis (Internship) - 3 months:** Project management and efficient communication
- **Jamie (Working Student) - Current:** Hands-on experience with advanced AI systems

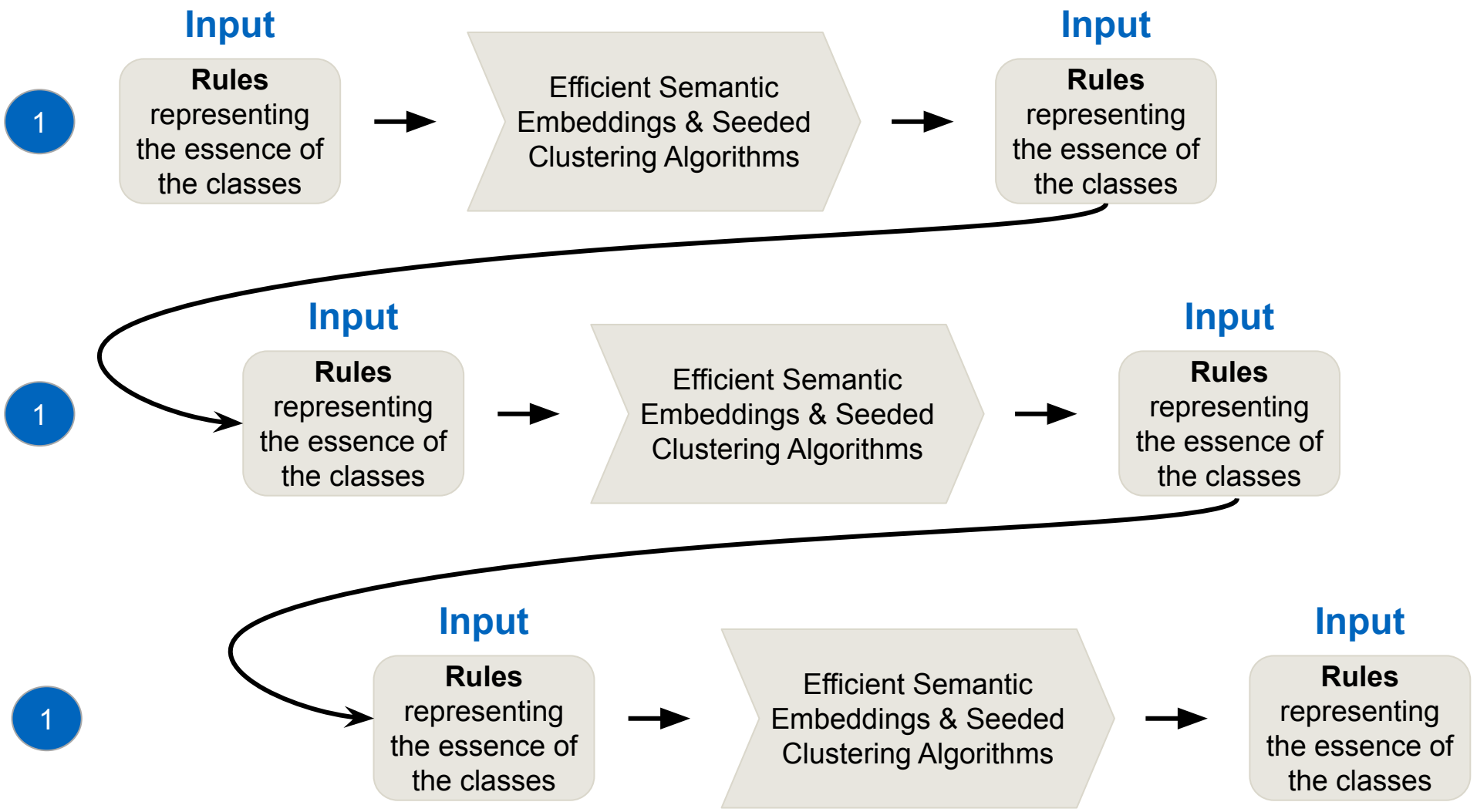
Tasks & Timeline

Official Start Date: 15th of January 2024

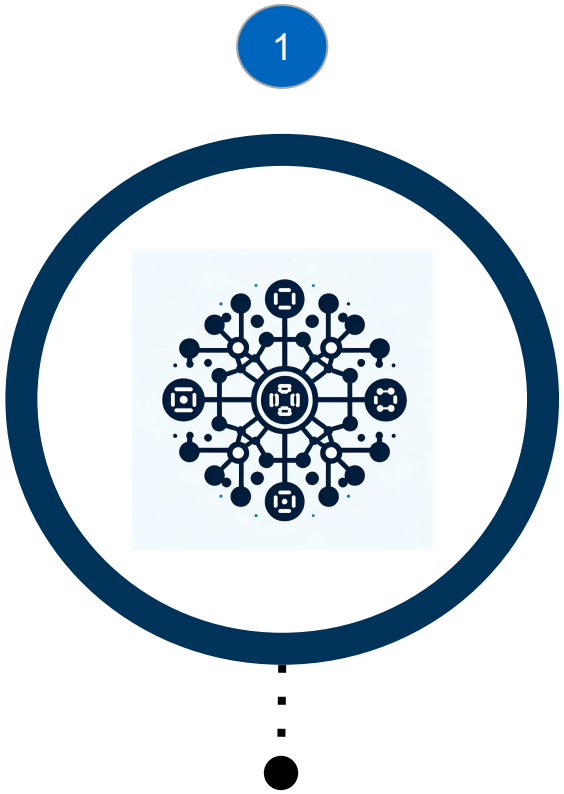
Official Submission Date: 15th of May 2024

Tasks	Timeline
Data inspection & analysis	Week 1
Research into efficient semantic embeddings & seeded clustering + Implementation	Week 2 - 4
Research into text ranking + Implementation	Week 5 - 7
Research into active learning & ensemble models + Implementation	Week 8 - 10
Writing of thesis & creation of final presentation	Week 11 - 16

Proposed Process Flow



Proposed Process Flow



Stick to the good sebis traditions

- Provide action links at the bottom of the slide to guide the audience to our web pages or publications (see below). (Select the text, press CTRL-K)

- Use a file name according to our sebis conventions which helps us and our audience to find the file of your presentation on our web site with Google search:
  - YYMMDD Author Short Title
  - Include this string in the footer (Einfügen -> Kopf- und Fusszeile -> Fusszeile)
  - The unusual date format simplifies the search for the latest version of a slide in an alphabetical directory listing (Dropbox, Explorer, Tricia, Sky-Drive)

[Ha13g] Hauder, M., Roth, S., Matthes, F.: Current Tool support for Metrics in Enterprise Architecture Management

For more information visit BEAMS , EAM Pattern Catalog and EAM KPI Catalog (http://wwwmatthes.in.tum.de)

Use the sebis visual language (shapes, fonts, colors, sizes)