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

Markus Löhde, B.Sc. Informatics

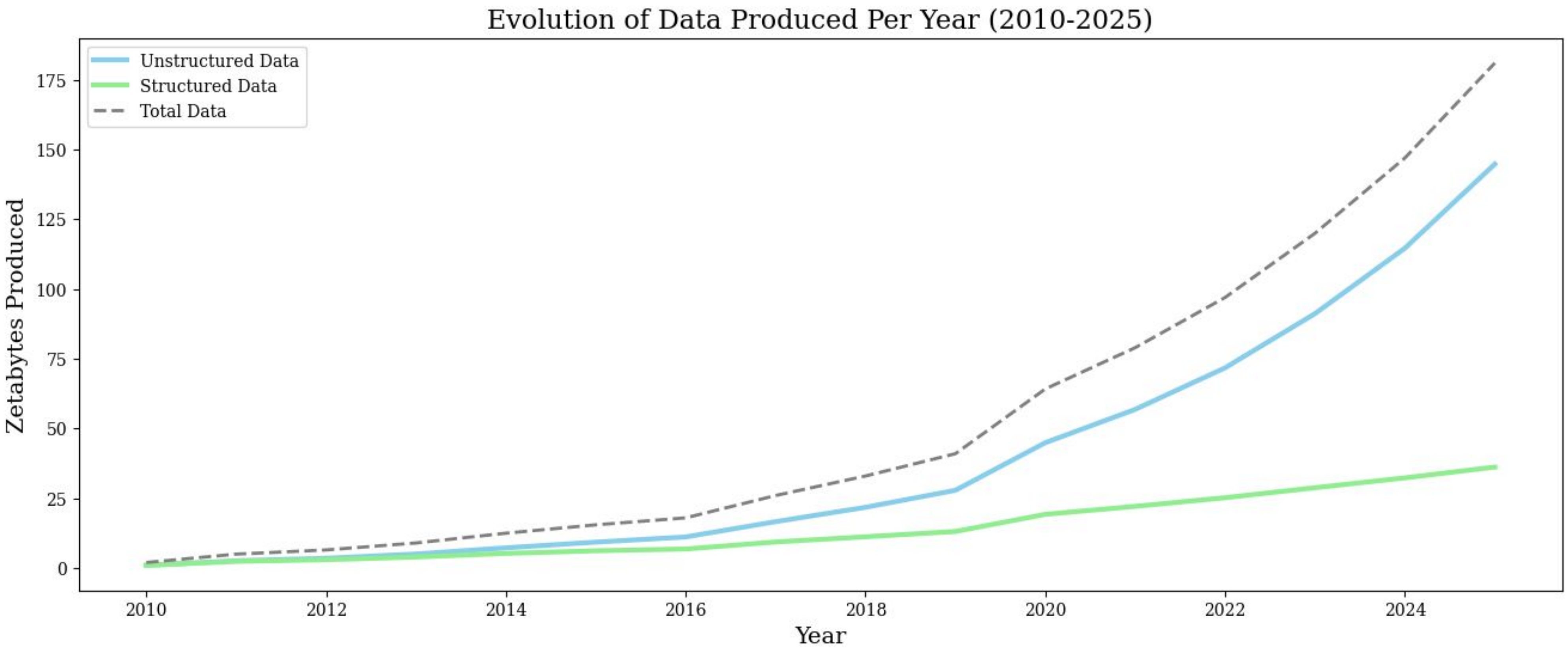
27.05.2024, Final Presentation

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

Evolution of Data Creation

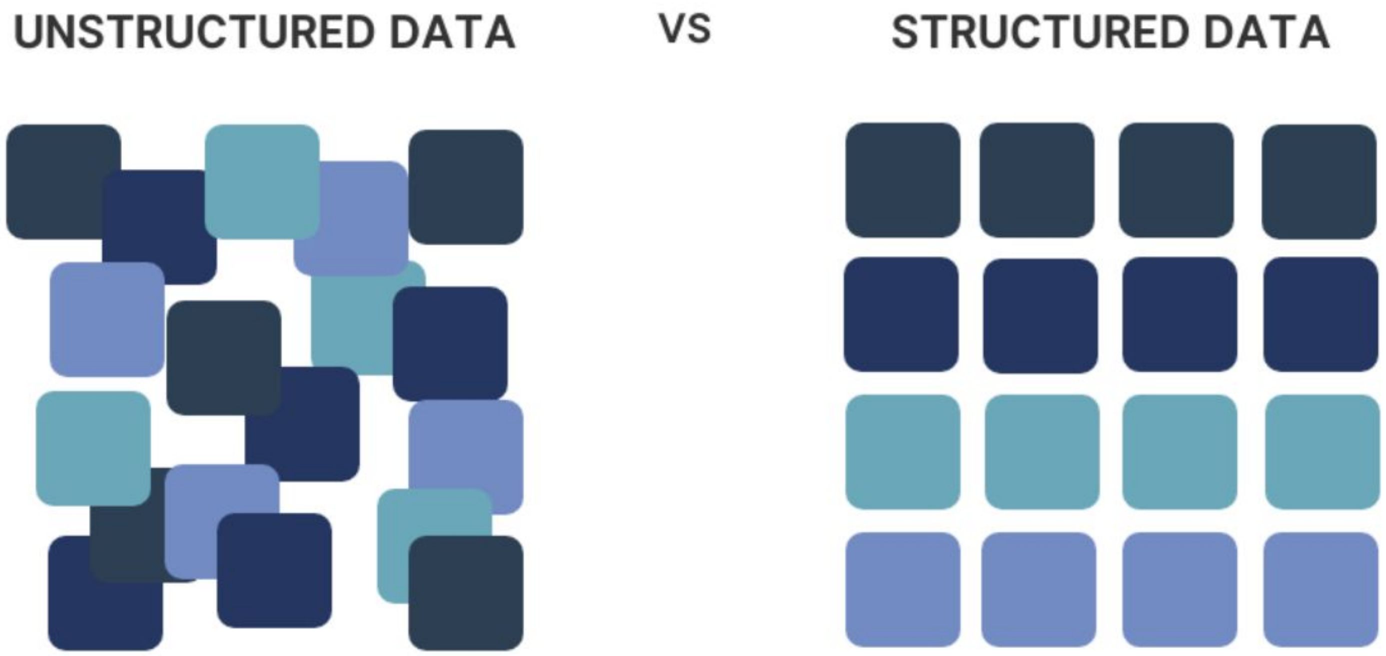
- With data creation increasing exponentially, we expect to produce 150 zetabytes globally in 2024. [1]
- However ~80% of that data will be unstructured! [7]



[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

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

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

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

Keyword Extraction → Context Windows → Extrapo-lation

“Betreiben eines Bauernhofs”

Output

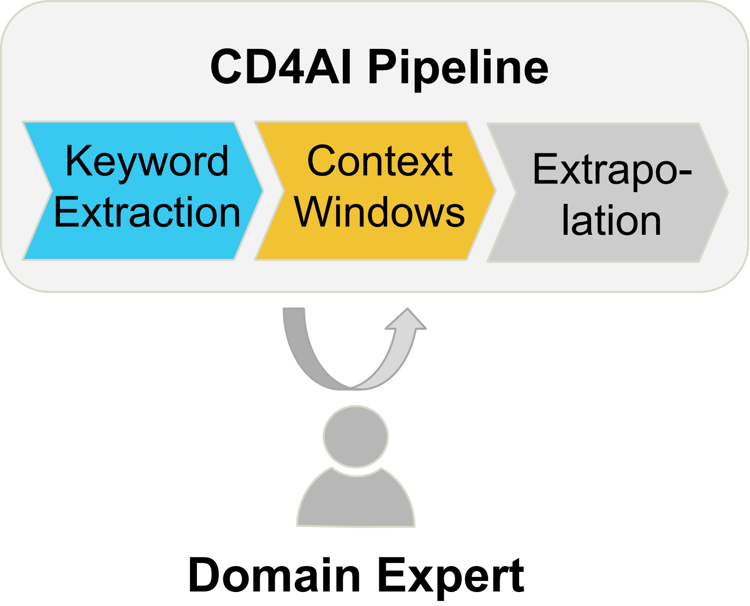
CreateData4AI (CD4AI)

- 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
...	...



Output

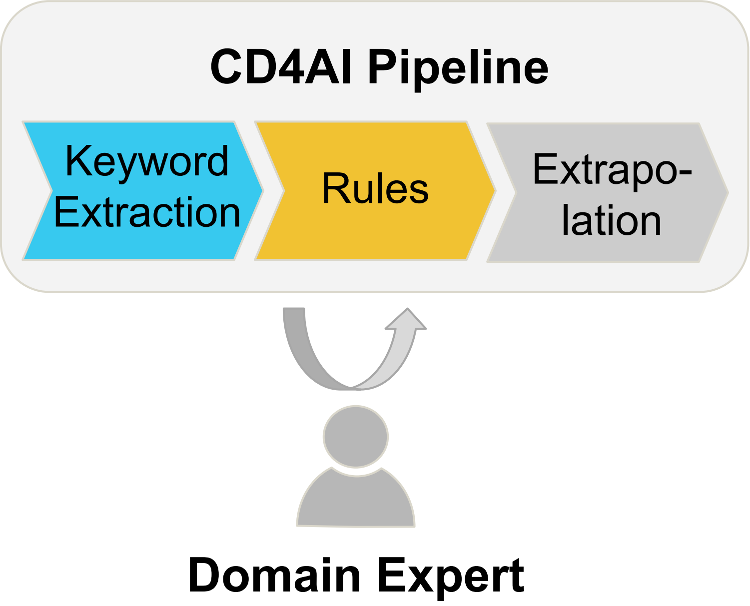
CreateData4AI (CD4AI)

- 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
...	...



Output

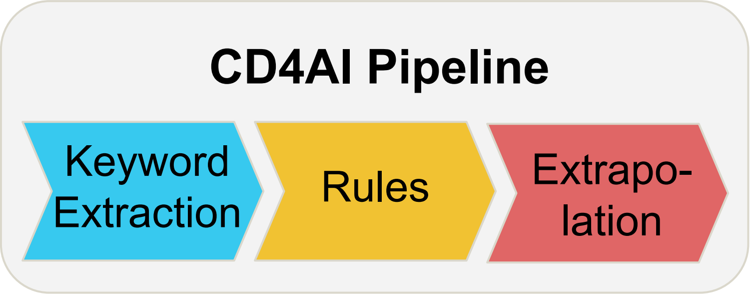
CreateData4AI (CD4AI)

- 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
...	...



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)

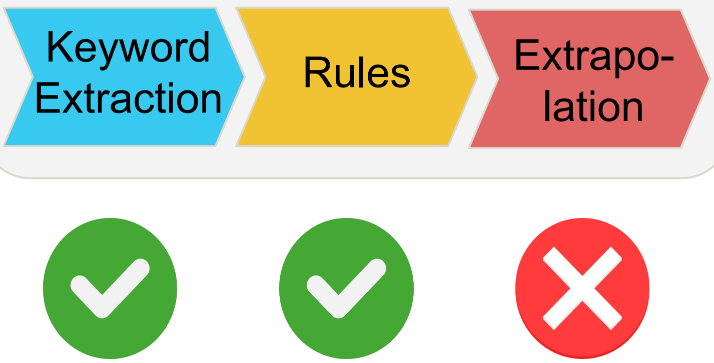
- In the CD4AI project we aim to develop a data-annotation pipeline with a human in the loop
- Our pilot project deals with a 3 million row dataset from the german trade register that details the purpose of companies
- The companies need to be categorized into 21 classes, corresponding to the 21 economic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

Research Questions

Main Research Question:

How can current state-of-the-art NLP techniques be used for a multilabel classification of large, domain-specific text corpora?

Research Questions

Main Research Question:

How can current state-of-the-art NLP techniques be used for a multilabel classification of large, domain-specific text corpora?

Supporting Research Questions:

	Supporting Research Questions	Answer Approach
1	What are the state-of-the-art approaches for a multilabel classification of large, domain-specific text corpora?	**Literature review**

Research Questions

Main Research Question:

How can current state-of-the-art NLP techniques be used for a multilabel classification of large, domain-specific text corpora?

Supporting Research Questions:

Answer Approach:

	Supporting Research Question	Answer Approach
1	What are the state-of-the-art approaches for a multilabel classification of large, domain-specific text corpora?	**Literature review**
2	What is the most efficient and accurate approach for leveraging context-specific class archetypes for a multilabel classification of large, domain-specific text corpora?	**Exploration & Experimentation**

Research Questions

Main Research Question:

How can current state-of-the-art NLP techniques be used for a multilabel classification of large, domain-specific text corpora?

Supporting Research Questions:

	Supporting Research Questions	Answer Approach
1	What are the state-of-the-art approaches for a multilabel classification of large, domain-specific text corpora?	**Literature review**
2	What is the most efficient and accurate approach for leveraging context-specific class archetypes for a multilabel classification of large, domain-specific text corpora?	**Exploration & Experimentation**
3	How can the efficiency and accuracy of a system designed to annotate large, domain-specific text corpora be evaluated?	**Research into popular metrics**

Outline

RQ1: Literature Review

- At its core the extrapolation step of the CD4AI pipeline is a multilabel text classification task.
- The following table summarises the three major research fields for this kind of task and rates the immediate applicability of the state-of-the-art methods of each field to the extrapolation step of CD4AI.

RQ1: Literature Review

- At its core the extrapolation step of the CD4AI pipeline is a multilabel text classification task.
- The following table summarises the three major research fields for this kind of task and rates the immediate applicability of the state-of-the-art methods of each field to the extrapolation step of CD4AI.

Research Field	**Description**	**State-of-the-Art**	**Applicability**
Zero-Shot Classification	• Zero-shot classification deals with scenarios in which no labels for the dataset are available. • As a result, the models often aim to possess a general understanding of human language.	• NLI-based models like facebook/bart-large-mnli [8] • LLMs like GPT-4	

RQ1: Literature Review

- At its core the extrapolation step of the CD4AI pipeline is a multilabel text classification task.
- The following table summarises the three major research fields for this kind of task and rates the immediate applicability of the state-of-the-art methods of each field to the extrapolation step of CD4AI.

Research Field	**Description**	**State-of-the-Art**	**Applicability**
Zero-Shot Classification	• Zero-shot classification deals with scenarios in which no labels for the dataset are available. • As a result, the models often aim to possess a general understanding of human language.	• NLI-based models like facebook/bart-large-mnli [8] • LLMs like GPT-4	
Weakly-Supervised Classification	• Weakly-supervised classification deals with scenarios where only imprecise labels are available. • Examples would be label descriptions or keywords.	• Approaches based on pseudo-document generation and self-training [9]	

RQ1: Literature Review

- At its core the extrapolation step of the CD4AI pipeline is a multilabel text classification task.
- The following table summarises the three major research fields for this kind of task and rates the immediate applicability of the state-of-the-art methods of each field to the extrapolation step of CD4AI.

Research Field	Description	State-of-the-Art	Applicability
Zero-Shot Classification	• Zero-shot classification deals with scenarios in which no labels for the dataset are available. • As a result, the models often aim to possess a general understanding of human language.	• NLI-based models like facebook/bart-large-mnli [8] • LLMs like GPT-4	
Weakly-Supervised Classification	• Weakly-supervised classification deals with scenarios where only imprecise labels are available. • Examples would be label descriptions or keywords.	• Approaches based on pseudo-document generation and self-training [9]	
Few-Shot Classification	• Few-shot classification deals with scenarios in which a few high-quality, labeled examples are available. • The labeling process often requires human effort.	• Sentence Transformers fine-tuned via SetFit [5]	

Outline

RQ2: Methodology

- The following algorithm serves as a blueprint for all the methods that we developed in our thesis.

Algorithm 3 Abstract Algorithm

```
procedure AbstractSimilarityMatching(docs, rules, k, m)
    result ← []
    for doc in docs do
        topMRules ← rankRulesBySimilarity(doc, rules, m)
        topKClasses ← getTopKClasses(topMRules, k)
        append topKClasses to result
    end for
    return result
end procedure
```

- `rankRulesBySimilarity` orders all rules according to a similarity criterion that is specific to each method.
- `m` determines determines the number of top-matching rules to be determined by `rankRulesBySimilarity`.
- `getTopKClasses` returns the `k` classes with the highest frequency among the top `m` rules.

RQ2: Methodology

- The following algorithm serves as a blueprint for all the methods that we developed in our thesis.

Algorithm 3 Abstract Algorithm

procedure AbstractSimilarityMatching(*docs*, *rules*, *k*, *m*)
    result $\leftarrow$ []
    for *doc* **in** *docs* **do**
        topMRules $\leftarrow$ **rankRulesBySimilarity**(*doc*, *rules*, *m*)
        topKClasses $\leftarrow$ **getTopKClasses**(*topMRules*, *k*)
        append *topKClasses* **to** *result*
    end for
    return *result*
end procedure

- `rankRulesBySimilarity` orders all rules according to a similarity criterion that is specific to each method.
- `m` determines determines the number of top-matching rules to be determined by `rankRulesBySimilarity`.
- `getTopKClasses` returns the `k` classes with the highest frequency among the top `m` rules.

RQ2: Methodology

- The following algorithm serves as a blueprint for all the methods that we developed in our thesis.

Algorithm 3 Abstract Algorithm

```
procedure AbstractSimilarityMatching(docs, rules, k, m)
    result ← []
    for doc in docs do
        topMRules ← rankRulesBySimilarity(doc, rules, m)
        topKClasses ← getTopKClasses(topMRules, k)
        append topKClasses to result
    end for
    return result
end procedure
```

- `rankRulesBySimilarity` orders all rules according to a similarity criterion that is specific to each method.
- `m` determines determines the number of top-matching rules to be determined by `rankRulesBySimilarity`.
- `getTopKClasses` returns the `k` classes with the highest frequency among the top `m` rules.

RQ2: Methodology

- The following algorithm serves as a blueprint for all the methods that we developed in our thesis.

Algorithm 3 Abstract Algorithm

```
procedure AbstractSimilarityMatching(docs, rules, k, m)
    result ← []
    for doc in docs do
        topMRules ← rankRulesBySimilarity(doc, rules, m)
        topKClasses ← getTopKClasses(topMRules, k)
        append topKClasses to result
    end for
    return result
end procedure
```

- `rankRulesBySimilarity` orders all rules according to a similarity criterion that is specific to each method.
- `m` determines determines the number of top-matching rules to be determined by `rankRulesBySimilarity`.
- `getTopKClasses` returns the *k* classes with the highest frequency among the top *m* rules.

RQ2: Methodology

- The following algorithm serves as a blueprint for all the methods that we developed in our thesis.

Algorithm 3 Abstract Algorithm

procedure AbstractSimilarityMatching(*docs*, *rules*, *k*, *m*)
    result ← []
    for *doc* **in** *docs* **do**
        topMRules ← **rankRulesBySimilarity**(*doc*, *rules*, *m*)
        topKClasses ← **getTopKClasses**(*topMRules*, *k*)
        append *topKClasses* **to** *result*
    end for
    return *result*
end procedure

- `rankRulesBySimilarity` orders all rules according to a similarity criterion that is specific to each method.
- `m` determines determines the number of top-matching rules to be determined by `rankRulesBySimilarity`.
- `getTopKClasses` returns the `k` classes with the highest frequency among the top `m` rules.

RQ2: Methodology

- The following algorithm serves as a blueprint for all the methods that we developed in our thesis.

Algorithm 3 Abstract Algorithm

```
procedure AbstractSimilarityMatching(docs, rules, k, m)
    result ← []
    for doc in docs do
        topMRules ← rankRulesBySimilarity(doc, rules, m)
        topKClasses ← getTopKClasses(topMRules, k)
        append topKClasses to result
    end for
    return result
end procedure
```

- `rankRulesBySimilarity` orders all rules according to a similarity criterion that is specific to each method.
- `m` determines determines the number of top-matching rules to be determined by `rankRulesBySimilarity`.
- `getTopKClasses` returns the `k` classes with the highest frequency among the top `m` rules.

RQ2: Methodology

- The following algorithm serves as a blueprint for all the methods that we developed in our thesis.

Algorithm 3 Abstract Algorithm

procedure AbstractSimilarityMatching(*docs*, *rules*, *k*, *m*)
    result ← []
    for *doc* **in** *docs* **do**
        topMRules ← **rankRulesBySimilarity**(*doc*, *rules*, *m*)
        topKClasses ← **getTopKClasses**(*topMRules*, *k*)
        append *topKClasses* **to** *result*
    end for
    return *result*
end procedure

- `rankRulesBySimilarity` orders all rules according to a similarity criterion that is specific to each method.
- `m` determines determines the number of top-matching rules to be determined by `rankRulesBySimilarity`.
- `getTopKClasses` returns the `k` classes with the highest frequency among the top `m` rules.

RQ2: Methodology

- The following algorithm serves as a blueprint for all the methods that we developed in our thesis.

Algorithm 3 Abstract Algorithm

```
procedure AbstractSimilarityMatching(docs, rules, k, m)
    result ← []
    for doc in docs do
        topMRules ← rankRulesBySimilarity(doc, rules, m)
        topKClasses ← getTopKClasses(topMRules, k)
        append topKClasses to result
    end for
    return result
end procedure
```

- `rankRulesBySimilarity` orders all rules according to a similarity criterion that is specific to each method.
- *m* determines determines the number of top-matching rules to be determined by `rankRulesBySimilarity`.
- `getTopKClasses` returns the *k* classes with the highest frequency among the top *m* rules.

RQ2: Methodology

- The following algorithm serves as a blueprint for all the methods that we developed in our thesis.

Algorithm 3 Abstract Algorithm

```
procedure AbstractSimilarityMatching(docs, rules, k, m)
    result ← []
    for doc in docs do
        topMRules ← rankRulesBySimilarity(doc, rules, m)
        topKClasses ← getTopKClasses(topMRules, k)
        append topKClasses to result
    end for
    return result
end procedure
```

- `rankRulesBySimilarity` orders all rules according to a similarity criterion that is specific to each method.
- *m* determines determines the number of top-matching rules to be determined by `rankRulesBySimilarity`.
- `getTopKClasses` returns the *k* classes with the highest frequency among the top *m* rules.

RQ2: Methodology

- The following algorithm serves as a blueprint for all the methods that we developed in our thesis.

Algorithm 3 Abstract Algorithm

```
procedure AbstractSimilarityMatching(docs, rules, k, m)
    result ← []
    for doc in docs do
        topMRules ← rankRulesBySimilarity(doc, rules, m)
        topKClasses ← getTopKClasses(topMRules, k)
        append topKClasses to result
    end for
    return result
end procedure
```

- `rankRulesBySimilarity` orders all rules according to a similarity criterion that is specific to each method.
- `m` determines determines the number of top-matching rules to be determined by `rankRulesBySimilarity`.
- `getTopKClasses` returns the `k` classes with the highest frequency among the top `m` rules.

RQ2: Methodology

- The following algorithm serves as a blueprint for all the methods that we developed in our thesis.

Algorithm 3 Abstract Algorithm

```
procedure AbstractSimilarityMatching(docs, rules, k, m)
    result ← []
    for doc in docs do
        topMRules ← rankRulesBySimilarity(doc, rules, m)
        topKClasses ← getTopKClasses(topMRules, k)
        append topKClasses to result
    end for
    return result
end procedure
```

- `rankRulesBySimilarity` orders all rules according to a similarity criterion that is specific to each method.
- `m` determines determines the number of top-matching rules to be determined by `rankRulesBySimilarity`.
- `getTopKClasses` returns the `k` classes with the highest frequency among the top `m` rules.

Exact String Matching

- Here the similarity criterion is whether the string representation of a rule can be found as an exact substring inside of document. Therefore, the similarity is binary: a rule either matches a document completely or not at all.
- The following example to showcases how the abstract matching algorithm works in the case of Exact String Matching:

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'

Exact String Matching - Example

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']
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'

Rule	Class	Similarity

Exact String Matching - Example

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']
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'

Rule	Class	Similarity

Exact String Matching - Example

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']
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'

Rule	Class	Similarity

Exact String Matching - Example

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']
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'

Rule	Class	Similarity
‘betrieb einer spedition’	H	100%

Exact String Matching - Example

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']
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb', 'verwaltung und geschäftsführung', 'kaufmännische beratung']
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'

Rule	Class	Similarity
‘betrieb einer spedition’	H	100%
‘der betrieb’	M	100%

Exact String Matching - Example

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']
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'

Rule	Class	Similarity
'betrieb einer spedition'	H	100%
'der betrieb'	M	100%

Exact String Matching - Example

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']
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'

Rule	Class	Similarity
'betrieb einer spedition'	H	100%
'der betrieb'	M	100%

Exact String Matching - Example

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']
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'

Rule	Class	Similarity
'betrieb einer spedition'	H	100%
'der betrieb'	M	100%

Exact String Matching - Example

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']
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'

Rule	Class	Similarity
'betrieb einer spedition'	H	100%
'der betrieb'	M	100%

Exact String Matching - Example

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']
M	Erbringung von wirtschaftlichen und technischen Dienstleistungen	['der betrieb' , 'verwaltung und geschäftsführung', 'kaufmännische beratung']
I	Gastgewerbe	['Hotellerie und Touristik', 'Gastronomische Einrichtungen'

Rule	Class	Similarity
‘betrieb einer spedition’	H	100%
‘der betrieb’	M	100%

`getTopKClasses(k=2)` → [H, M]

Fuzzy String Matching

- In this method, the similarity between a rule and a document is based on the Levenshtein distance, which calculates the minimum number of insertions, deletions, and substitutions needed to convert one string into the other.
- Specifically, we utilize the function `partial_token_sort_ratio` from the python library `thefuzz` which performs three key steps encoded its name [10]:
  1. `partial`: The function takes the shorter string (the rule) as a reference and compares it to all substrings of the longer string (the document).
  2. `token_sort`: The function also sorts the tokens of both strings before comparing them, making the order of tokens irrelevant.
  3. `ratio`: Finally, the Levenshtein similarity is computed for all sorted substrings, yielding a continuous similarity score.
- Example:

1. "The fuzzy wuzzy bear!"
2. "The wuzzy fuzzy bear"

`partial_token_sort_ratio`

100%

Vanilla Semantic Similarity Matching

- Syntactic methods face inherent scalability and accuracy issues because the same meaning can be expressed in many ways.
- Consequently, we explored semantic similarity measures using sentence transformers [11] and cosine similarity, as shown below.

Document

“Der Betrieb einer Spedition und eines Transportunternehmens”

Rule

“betrieb einer spedition”

Sentence Transformer Embedding Model

[0.28251907, 0.03204801, 0.00893406, 0.05167645, 0.20162104, -0.04518226, 0.18206343, 0.07122982, -0.37910348, -0.11333402, ..]

[-0.00282591, 0.11538339, 0.22487034, 0.0701141 , 0.185079 ,-0.06716876, 0.12154326, 0.10647674, -0.31949738, -0.12266784, ..]

cosine_similarity

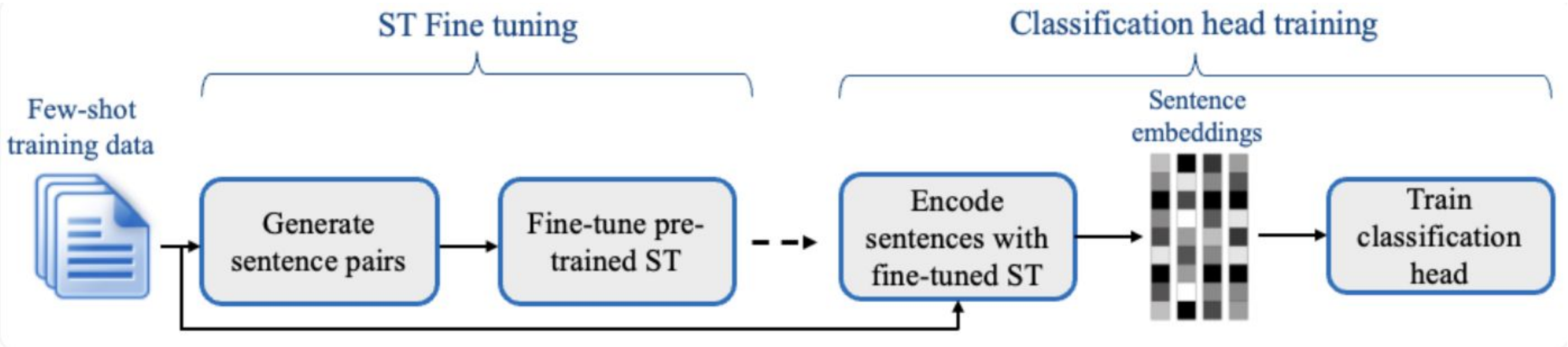
Fine-Tuned Semantic Similarity Matching

- The classification of two pieces of text as similar is heavily dependent on the context of the classification task.
- So, two pieces of text that are similar in a general sense are not automatically similar for our specific task of assigning companies to economic sectors.
- The following example illustrates this point: In our context (1) & (3) actually belong to the same economic sector (C) and (1) & (2) do not.

ID	Company Purpose
(1)	Das Herstellen und der Transport von Automobilen.
(2)	Das Vertreiben und die Reparatur von Automobilen.
(3)	Die Produktion von Ersatzteilen für schwere Maschinerie, wie zum Beispiel Traktoren, Fabrikroboter und Autos.
	General Similarity Scores
	Between (1) & (2): 0.8369
	Between (1) & (3): 0.4589

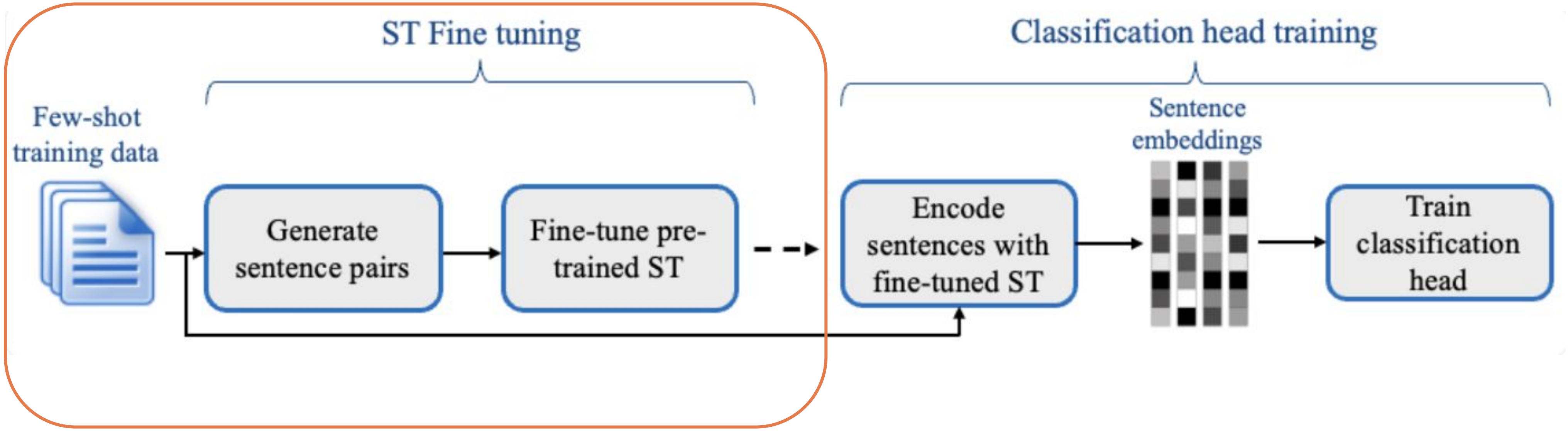
Fine-Tuned Semantic Similarity Matching

- In order to tailor our embeddings to the data our specific classification task, we utilize SetFit [5].
- As discussed in the literature review, SetFit is a framework that allows us to fine-tune any sentence transformer to our specific dataset. For example, we would like the embeddings of (1) & (3) to get more similar and those of (1) & (2) to get more dissimilar.
- However, SetFit requires a few labeled examples for this fine-tuning, which we lack. Therefore, we instead use the rules as noisy approximations of labeled documents.



Fine-Tuned Semantic Similarity Matching

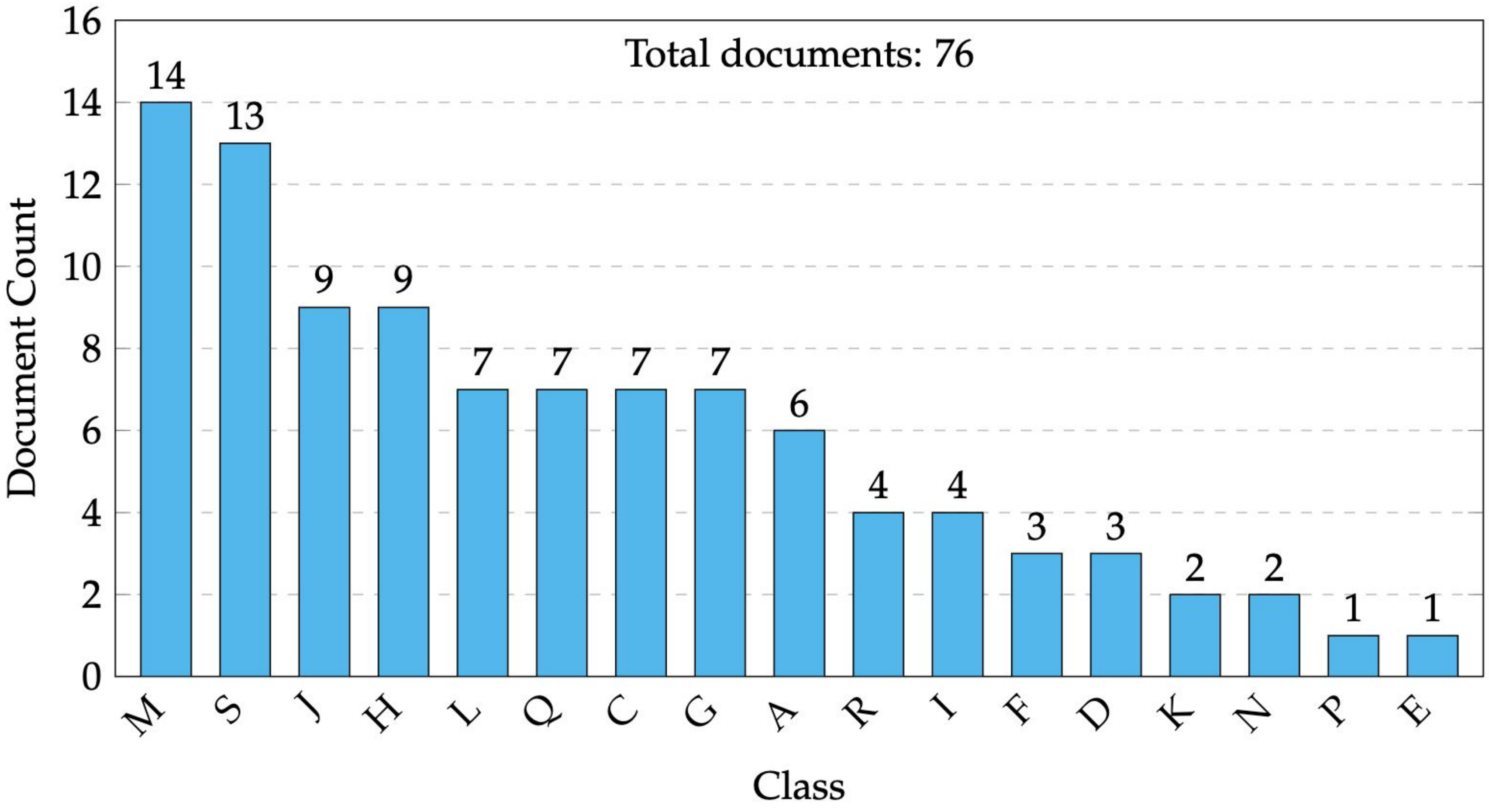
- In order to tailor our embeddings to the data our specific classification task, we utilize SetFit [5].
- As discussed in the literature review, SetFit is a framework that allows us to fine-tune any sentence transformer to our specific dataset. For example, we would like the embeddings of (1) & (3) to get more similar and those of (1) & (2) to get more dissimilar.
- However, SetFit requires a few labeled examples for this fine-tuning, which we lack. Therefore, we instead use the rules as noisy approximations of labeled documents.



Outline

Test Dataset

- To evaluate the effectiveness of our methods, we manually curated a test dataset consisting of 76 labeled documents. However, due to the long-tailed distribution of our data [12], the dataset is not entirely balanced, and certain classes, namely [O, U, T, B], are not represented at all.



Performance Metrics

- In order to quantify the accuracy and the efficiency of our methods we used the following metrics.

Metric	Description	Definition	Type
Precision	• Intuitively precision answers the question: *“If we predict that a document belongs to class X, how often are we correct?”*	$precision = \frac{tp}{tp + fp}$	
Recall	• Intuitively precision answers the question: *“Out of all the actual positive instances, how many did we correctly predict?”*	$recall = \frac{tp}{tp + fn}$	
F1-Score	• The f1-score integrates both precision and recall, which makes it a crucial measure for assessing the overall accuracy of our methods.	$f1 = 2 \times \frac{precision \times recall}{precision + recall}$	
Computation Time	• This metric signals the computation time per document on an L4 GPU in seconds.	$comp\_time = \frac{time(s)}{n\_documents}$	

Sets of Context Rules

- Syntactic and semantic methods use different sets of context rules due to their focus on distinct similarity aspects between rules and documents.
- The *Syntactic Rule Set* has at least 1500 rules per class and generally contains shorter rules, while the *Semantic Rule Set* has exactly 33 rules per class and generally contains longer, semantically richer rules.

Syntactic Rule Set
"von pflanzen", "von agrarprodukten und dienstleistungen", "urbanen und ländlichen raum", "fördermittel eu, bund, land", "landwirtschaftlichen betriebes", "insbesondere gartengeräte und der handel"

Semantic Rule Set
"Übernahme von Grundstücksbesitz, insbesondere Forstwirtschaft, und persönliche Haftung für Viehhand", "Übernahme der Geschäftsführung von Land & Gut Englberger GmbH durch Unbekannten", "Zusammenarbeit mit Bildungsinstitutionen, Verbänden, Kammern, Instituten, Kommunen und Unternehmen in Bere", "Verwaltung von Beteiligungen, Übernahme persönlicher Haftung in Handelsgesellschaften, Land- und Forst",

Results for our Methods

- The table below contains the best results with regard to the f1-score for all the methods described in our methodology.

Table 6.2.: The results for our own methods.

Method	Exact Str. Mat. [1]	Fuzzy Str. Mat. [2]	Vanilla Sem. Mat. [3]	SetFit Rule-Trained [4]
Precision	**0.4309**	0.4097	0.3925	0.4891
Recall	0.4343	0.4444	**0.5556**	0.3939
F1-score	0.3801	0.3790	**0.4435**	0.4093
Time (s) [GPU]	0.0922	0.5333	0.0931	**0.0577**

- We can see that the *"Vanilla Semantic Similarity Matching"* method achieves the highest f1-score by a significant margin. Notably, the fine-tuning the sentence transformer model on the rules decreases performance.

1. For Exact String Matching, we used n = 1000 rules per class.
2. For Fuzzy String Matching, we set m = 50 and n = 300.
3. For Vanilla Semantic Similarity Matching, we used the sentence transformer s = e5-large [13]
4. For SetFit Rule-Trained we also set s = e5-large
5. For all methods we set k = 3

Benchmark Results

- To assess the performance of our own methods, we have to compare them to other popular text classification techniques.
- As our benchmark models we chose GPT-4 and facebook/bart-large-mnli. Further, we fine-tuned a sentence transformer model on a manually curated training dataset via SetFit.

Metric	GPT-4 [1]	SetFit Example-Trained [2]	Bart-Large-MNLI
Precision	**0.5649**	0.4285	0.4059
Recall	0.6061	**0.7475**	0.5354
F1-score	**0.5480**	0.5306	0.4137
Time (s) [GPU]	0.6958	0.0231	0.4734

- In terms of f1-score GPT-4 achieves the best result. However, looking at the huge difference in computation time and marginal difference in f1-score the "SetFit Example-Trained" is most impressive.

1. For the computation time it is important to note that we called GPT-4 via the OpenAI API.
2. This time we used s = en-de-roberta [14] and included the logistic regression head for classification.

Benchmark Results

- To assess the performance of our own methods, we have to compare them to other popular text classification techniques.
- As our benchmark models we chose GPT-4 and facebook/bart-large-mnli. Further, we fine-tuned a sentence transformer model on a manually curated training dataset via SetFit.

Metric	GPT-4 [1]	SetFit Example-Trained [2]	Bart-Large-MNLI
Precision	**0.5649**	0.4285	0.4059
Recall	0.6061	**0.7475**	0.5354
F1-score	**0.5480**	0.5306	0.4137
Time (s) [GPU]	0.6958	0.0231	0.4734

- In terms of f1-score GPT-4 achieves the best result. However, looking at the huge difference in computation time and marginal difference in f1-score the "SetFit Example-Trained" is most impressive.

1. For the computation time it is important to note that we called GPT-4 via the OpenAI API.
2. This time we used s = en-de-roberta [14] and included the logistic regression head for classification.

Outline

Key Findings

- After analyzing both, the results of our methods and the results of the benchmark methods, we synthesized three key findings:

Key Findings

- After analyzing both, the results of our methods and the results of the benchmark methods, we synthesized three key findings:

1 Semantic methods are superior to Syntactic Methods

- In the results for our methods, the semantic methods outperformed the syntactic ones in both accuracy and efficiency.

Key Findings

- After analyzing both, the results of our methods and the results of the benchmark methods, we synthesized three key findings:

1 Semantic methods are superior to Syntactic Methods

- In the results for our methods, the semantic methods outperformed the syntactic ones in both accuracy and efficiency.

2 Fine-tuning sentence transformers on labeled examples yields excellent performance.

- The SetFit fine-tuned version of the semantic similarity method achieves higher accuracy compared to its vanilla counterpart. Further, it is close to GPT-4 with much higher efficiency.

Key Findings

- After analyzing both, the results of our methods and the results of the benchmark methods, we synthesized three key findings:

1 Semantic methods are superior to Syntactic Methods

- In the results for our methods, the semantic methods outperformed the syntactic ones in both accuracy and efficiency.

2 Fine-tuning sentence transformers on labeled examples yields excellent performance.

- The SetFit fine-tuned version of the semantic similarity method achieves higher accuracy compared to its vanilla counterpart. Further, it is close to GPT-4 with much higher efficiency.

3 Our current Semantic Rule Set inadequately approximates labeled examples

- Fine-tuning embedding models on labeled examples results in significantly higher accuracy compared to fine-tuning on the Semantic Rule Set.

Future Work

- We believe our thesis laid a solid foundation for future research into the extrapolation step of CD4AI.
- For further research, we make the following recommendations.

Future Work

- We believe our thesis laid a solid foundation for future research into the extrapolation step of CD4AI.
- For further research, we make the following recommendations.

Creating a more balanced test dataset.

As discussed, the current test dataset has a long-tailed distribution and only contains one or two documents for some of the classes.

Future Work

- We believe our thesis laid a solid foundation for future research into the extrapolation step of CD4AI.
- For further research, we make the following recommendations.

Creating a more balanced test dataset.

As discussed, the current test dataset has a long-tailed distribution and only contains one or two documents for some of the classes.

Testing on data from different domains.

Testing the performance of the methods on data from various domains will provide a more comprehensive overview of each method's effectiveness.

Future Work

- We believe our thesis laid a solid foundation for future research into the extrapolation step of CD4AI.
- For further research, we make the following recommendations.

Creating a more balanced test dataset.

As discussed, the current test dataset has a long-tailed distribution and only contains one or two documents for some of the classes.

Testing on data from different domains.

Testing the performance of the methods on data from various domains will provide a more comprehensive overview of each method's effectiveness.

Introducing thresholds for class prediction.

Always trying to predict k = 3 classes often reduces precision, as most documents have fewer correct classes. Using confidence thresholds could improve this.

Future Work

- We believe our thesis laid a solid foundation for future research into the extrapolation step of CD4AI.
- For further research, we make the following recommendations.

Creating a more balanced test dataset.

As discussed, the current test dataset has a long-tailed distribution and only contains one or two documents for some of the classes.

Testing on data from different domains.

Testing the performance of the methods on data from various domains will provide a more comprehensive overview of each method's effectiveness.

Introducing thresholds for class prediction.

Always trying to predict k = 3 classes often reduces precision, as most documents have fewer correct classes. Using confidence thresholds could improve this.

Adapting the approach by Meng et al.

The idea of pseudo-document generation and subsequent self-training sounds very promising to us. Future research would have to figure out how to use rules as seed knowledge.

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

[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

[7] The Unseen Data Conundrum

[8] facebook/bart-large-mnli · Hugging Face

[9] Weakly-Supervised-Neural-Text-Classification

[10] GitHub - seatgeek/thefuzz: Fuzzy String Matching in Python

References - 2

[11] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Conference on Empirical Methods in Natural Language Processing*.

[12] Unternehmen nach Wirtschaftsabschnitten im Berichtsjahr 2021 - Statistisches Bundesamt

[13] intfloat/multilingual-e5-large-instruct · Hugging Face

[14] T-Systems-onsite/cross-en-de-roberta-sentence-transformer