Towards Scalable Domain-Specific Document Annotation:
A Semantic Archetype-Driven Framework

Moritz Steigerwald

May 12th 2025, Kickoff Master's Thesis

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 & Motivation

Research Questions

Methodology

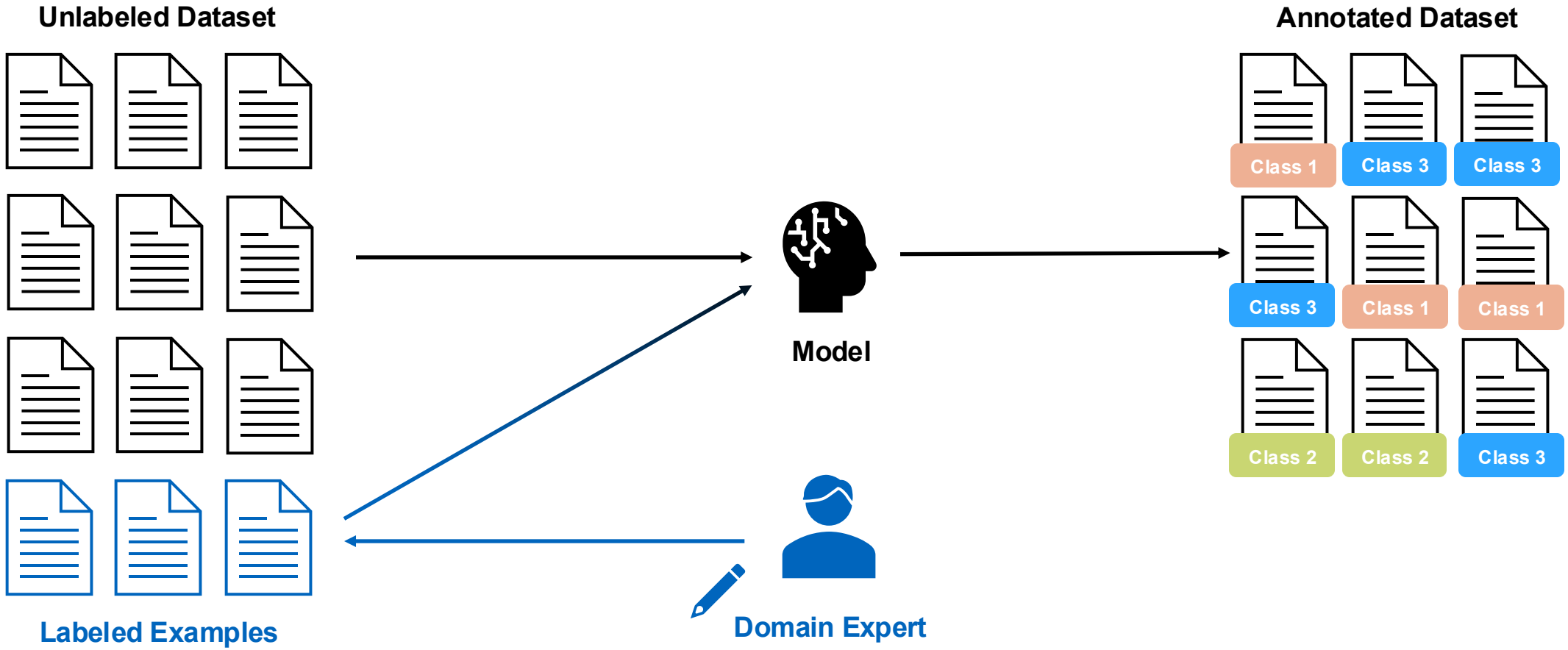
Evaluation

Expected Outcomes

Project Context

CreateData4AI: Leveraging Domain Knowledge and Context Rules to Transform Large-Scale Unstructured Text Corpora into Structured and Annotated Datasets

Introduction & Motivation



Research Gap & Challenges

Current Limitations:

- Manual annotation is **costly** and annotated data is **scarce**
- LLMs lack **domain specificity**.
- Existing classifiers **struggle with lengthy texts** and evolving domains.

Key Challenges:

- **Scalability** for large documents and datasets
- Bridging expert rules with **automated classification**
- Ensuring **adaptability** to domain shifts

Status Quo Step 0: Collecting the Text Corpus

[…] In many academic settings, modern supervised machine learning models rely heavily on large annotated text corpora. However, some researchers argue that unsupervised methods can outperform supervised techniques when domain-specific data is limited. In this paper, we develop a transformer-based architecture to evaluate advanced keyword extraction approaches.

Our approach compares dense embeddings from a pretrained language model with statistical features to identify relevant terminology in academic manuscripts.

We find that domain-specific metadata, when used effectively, helps refine context windows for more accurate classification. […]

Raw Text: No annotations.

- We have a large corpus of **unstructured text**
- No labels, **no annotations**
- This excerpt is one example

Status Quo Step 1: Keyword Extraction

[…] In many academic settings, modern **supervised machine learning models** rely heavily on **large annotated text corpora**. However, some researchers argue that unsupervised methods can outperform supervised techniques when domain-specific data is limited. In this paper, we develop a **transformer-based** architecture to evaluate advanced keyword extraction approaches.

Our approach compares **dense embeddings** from a **pretrained language model** with statistical features to identify relevant terminology in academic manuscripts.

We find that domain-specific **metadata**, when used effectively, helps refine context windows for more accurate **classification**. […]

Text with Extracted Keywords

- Apply Keyword Extraction Algorithm (e.g., KeyBERT, RAKE).
- Extracted keywords from this snippet
- Next: We gather “context windows”

Status Quo Step 2: Context Windows

In many academic settings, modern **supervised machine learning models** rely heavily on **large annotated text corpora**.

0

In this paper, we develop a **transformer-based** architecture to evaluate advanced keyword extraction approaches.

1

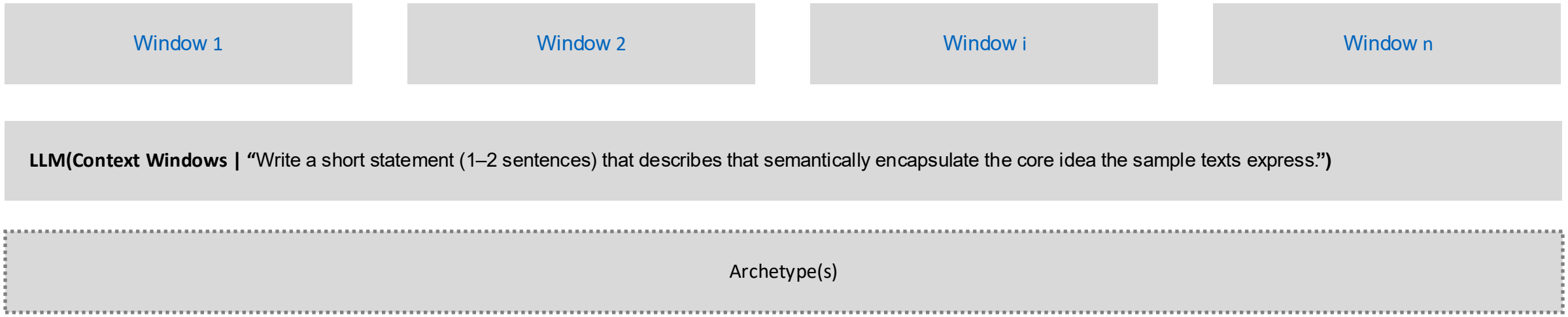
Our approach compares **dense embeddings** from a **pretrained language model** with statistical features to identify relevant terminology in academic manuscripts.

i

Extracted Context Windows

- These windows let us see how each keyword is used in the text
- Next: Derive archetypes from repeated patterns

Status Quo Step 3: Deriving Archetypes from Context Windows



Semantic Archetypes

- Next: **?**

Extra – Example Doc

[…] Inevitably they insist there is no need to raise taxes to fund improvements in services. The Tories claim they can improve services AND cut taxes through £35bn efficiency savings, while Labour has offered £22bn savings but has yet to map out precise tax proposals, although there is little chance they will propose increases. In many ways the argument between the Lib Dems and the others over taxation and spending echo the sort of arguments that raged between Labour and the Tories in the 1980s and early 1990s. But, unlike the old Tory-Labour debate, he believes voters are ready to see "modest" tax increases on the well off in order to fund improvements in services. That is a view partly endorsed by recent polls suggesting people would rather have cash spent on public services than tax cuts.

Similarly there is a different tone to the Lib Dem approach to asylum and immigration, with Mr Kennedy stressing politicians should not "foment an artificial debate" about immigration and attacking Michael Howard's proposals for quotas. Once again, with the two other big parties singing similar songs on immigration, Mr Kennedy is stressing the different, more liberal approach of his party. Mr Kennedy was also in buoyant mood over his party's election chances, declaring the Tories were not going to be "significant players" in the poll. He repeated his pledge not to do post-election deals with either party after the election. Mr Kennedy went on to suggest the re-election of a Labour government with a small majority would amount to a "massive vote of no confidence " in Tony Blair's government. That suggests the Lib Dem leader believes he may well find himself in a powerful, even pivotal position in a vastly different House of Commons after the next election. It is a dream the third party has dreamed many times before.

* Excerpt from BBC News dataset (Label: Politics)

Extra – Example Keywords

[…] Inevitably they insist there is no need to raise taxes to fund improvements in services. The Tories claim they can improve services AND cut taxes through £35bn efficiency savings, while Labour has offered £22bn savings but has yet to map out precise tax proposals, although there is little chance they will propose increases. In many ways the argument between the Lib Dems and the others over taxation and spending echo the sort of arguments that raged between Labour and the Tories in the 1980s and early 1990s. But, unlike the old Tory-Labour debate, he believes voters are ready to see "modest" tax increases on the well off in order to fund improvements in services. That is a view partly endorsed by recent polls suggesting people would rather have cash spent on public services than tax cuts.

Similarly there is a different tone to the Lib Dem approach to asylum and immigration, with Mr Kennedy stressing politicians should not "foment an artificial debate" about immigration and attacking Michael Howard's proposals for quotas. Once again, with the two other big parties singing similar songs on immigration, Mr Kennedy is stressing the different, more liberal approach of his party. Mr Kennedy was also in buoyant mood over his party's election chances, declaring the Tories were not going to be "significant players" in the poll. He repeated his pledge not to do post-election deals with either party after the election. Mr Kennedy went on to suggest the re-election of a Labour government with a small majority would amount to a "massive vote of no confidence " in Tony Blair's government.

* Domain Expert sets “Seed Keywords”, set gets Expanded by CD4AI Pipeline

Extra – Example Context Windows

but has yet to map out precise tax proposals, although there is little

unlike the old Tory-Labour debate, he believes

. . .

stressing politicians should not "foment

his pledge not to do post-election deals with either party

* We get a lot of Context Windows from even more Documents

Extra – Example Archetypes

Context Windows

n

Labour Party's electoral strategies and campaign management Election coordinators' statements, promises for upcoming elections, Labour's election manifestos

m

Elections, democracy, and governance. Researching elections, understanding electoral devices, and promoting electoral reforms.

* LLM Reduces number of context windows n to m archetypes (n >> m)

What's missing?

We have a partially annotated corpus with:

- ✓ **1. Keywords**,
- ✓ **2. Context windows**,
- ✓ **3. Some semantic text descriptors (archetypes).**

But no **full** automation yet – the rest of the corpus is still unlabeled!

✗ 1. We **need** a **robust classifier** to

(a) use these rule texts and

(b) systematically annotate all remaining data.

✗ 2. We also need to **evaluate** how well this system performs (accuracy, domain adaptability, etc.).

Research Questions

RQ1 How can semantic archetypes based on context rules be leveraged to classify and annotate domain-specific documents?

RQ2 Can a reward-based feedback system boost both classification accuracy and archetype quality?

RQ3 How does this archetype-driven framework compare to supervised models and zero-shot LLMs in terms of accuracy and resource utilization across different domains?

Methodology

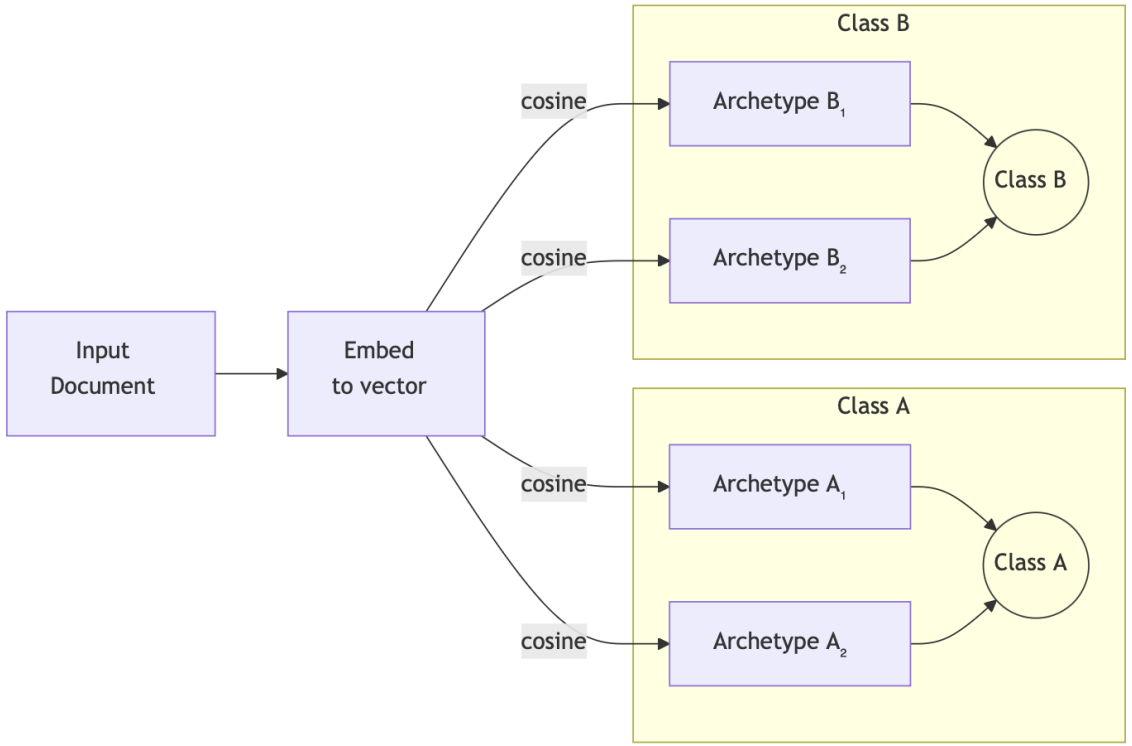
Partially Labeled Sets

- From domain experts or partial annotations
- Used to guide our classifier(s)

For Building a Classifier 4 Initial Methods Come to Mind:

1. Embedding Based Similarity
2. Archetype Anchored Contrastive Learning
3. Archetype Guided Self-Training Loop
4. RL & Feedback Loop

Methodology - Embedding Based Similarity

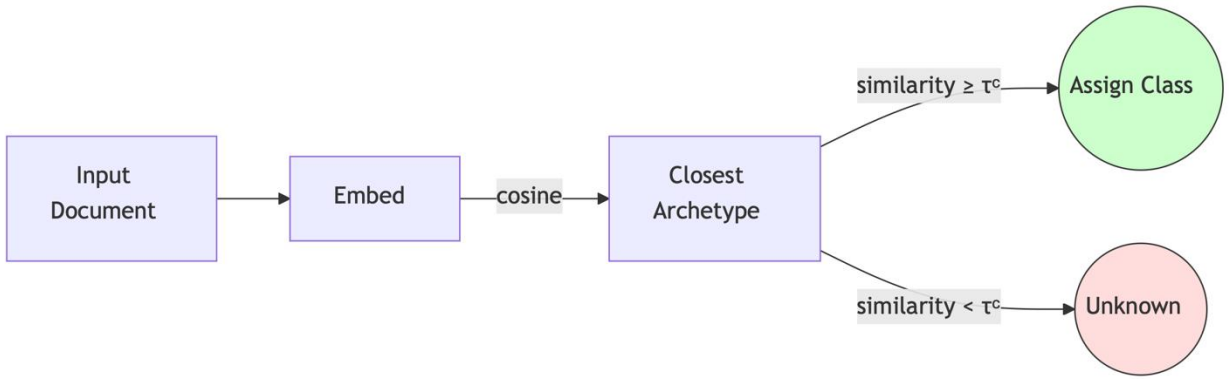


Nearest-Archetype (Rocchio) — *always picks the closest archetype.*

Problem: even out-of-scope docs get forced into one of the known classes.

Next → introduce similarity thresholds to allow an "unknown" outcome.

Methodology - Embedding Based Similarity cont'd I

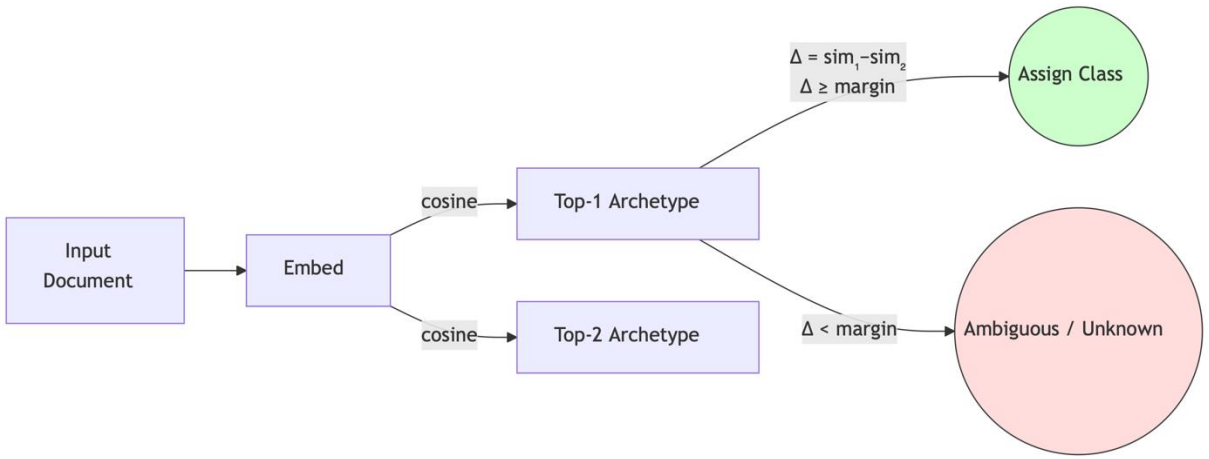


Thresholded Open-Set — *only assign a class if similarity ≥ its cutoff.*

Problem: documents that lie almost equally between two classes can still be mis-routed.

Next → add a cosine-margin check to catch ties and ambiguities.

Methodology - Embedding Based Similarity cont'd II

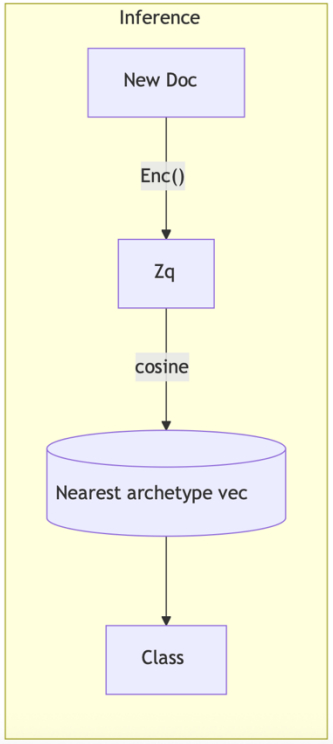


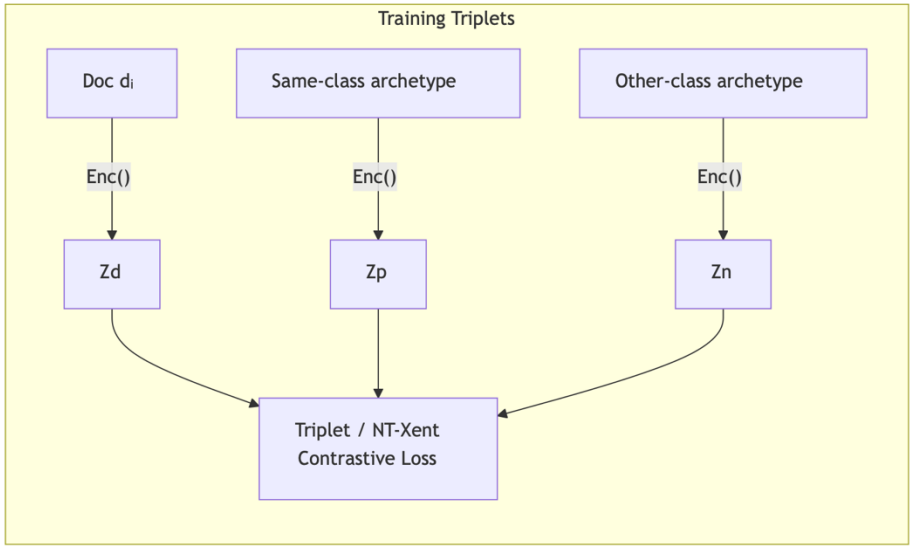
Cosine-Margin Guard — *require a clear gap (Δ) between top-1 and top-2 similarities before labeling.*

Problem: even with margins, embeddings may miss domain-specific semantics, limiting discriminative power.

Next → explore domain-adapted or supervised embedding strategies to capture those domain specifics

Methodology – Archetype-Anchored Contrastive Learning





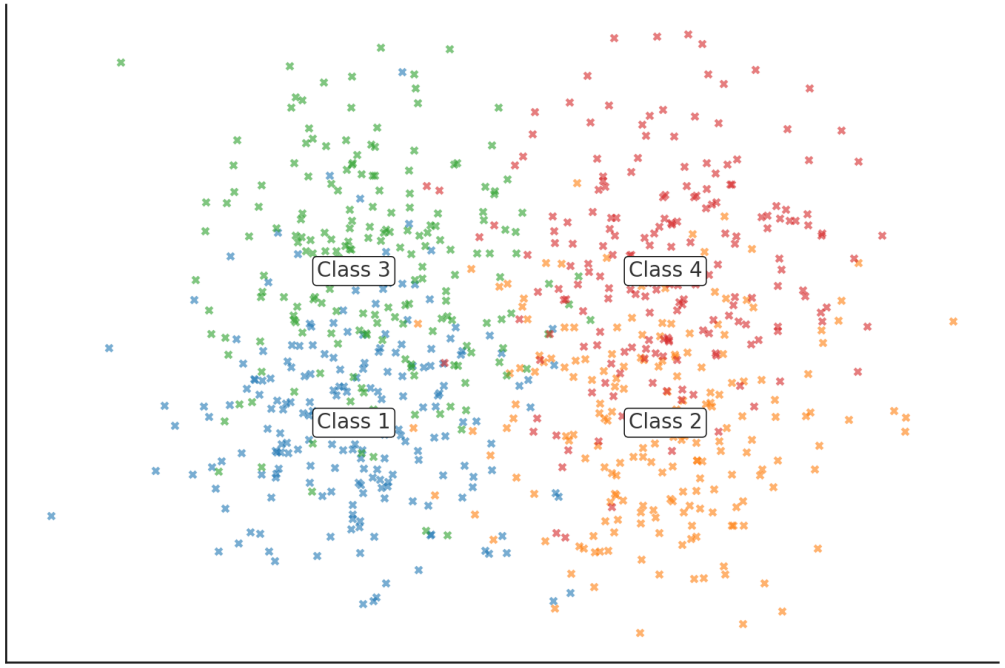
Contrastive Learning — train the encoder so documents are attracted to their own-class archetype embeddings and repelled from other-class archetypes

Problem: even after contrastive tuning, generic embeddings may still miss domain specifics and what are the correct thresholds?

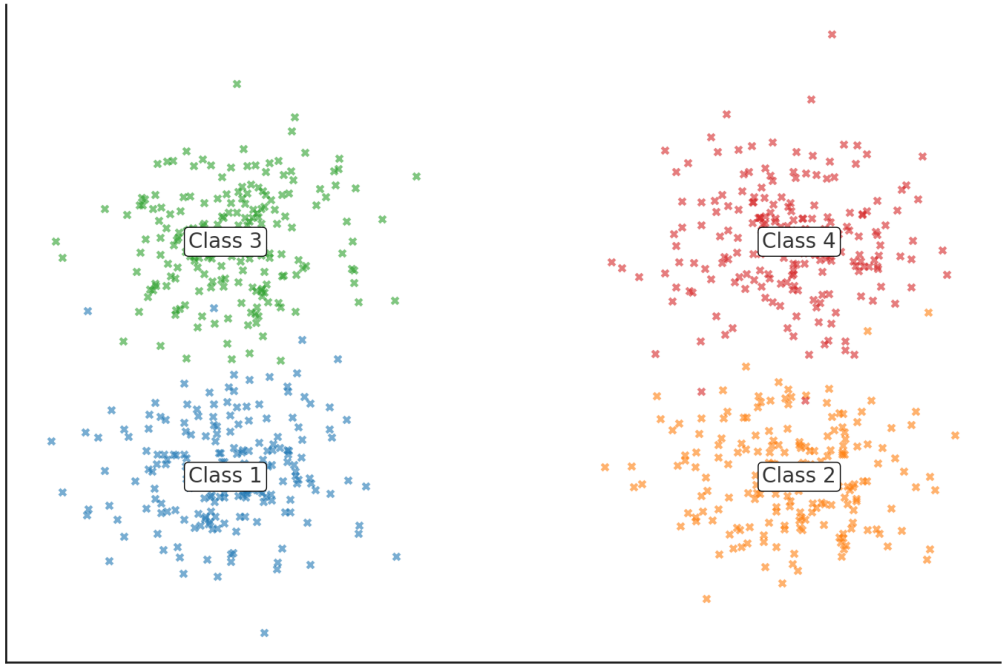
Next → explore fully guided training loop

Methodology – Archetype-Anchored Contrastive Learning cont'd

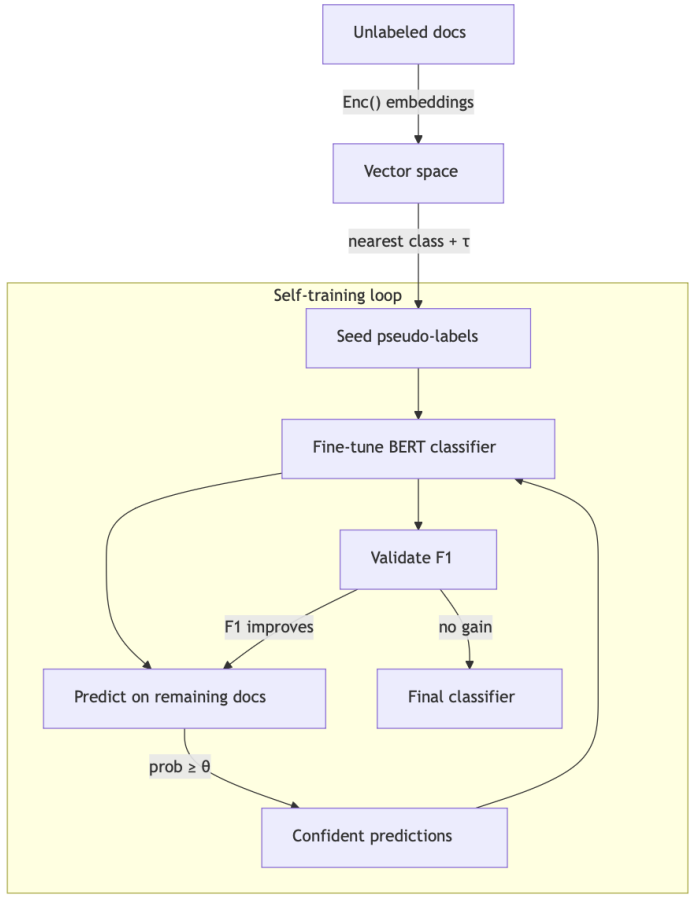
Before fine-tuning



After fine-tuning



Methodology – Archetype Guided Self-Training Loop



- **Seed stage:** embeddings + nearest-archetype → clean pseudo-labels (confidence τ).
- **Train stage:** fine-tune BERT on seed data (generic or embedding-tuned).
- **Self-train loop:** add high-probability predictions (threshold θ) each round.
- **Stop criterion:** stop when validation F1 plateaus ⇒ scientifically robust.
- **Result:** domain-adapted classifier without costly hand labels, leveraging embeddings only for the first “push.”

Methodology – RL & Feedback Loop

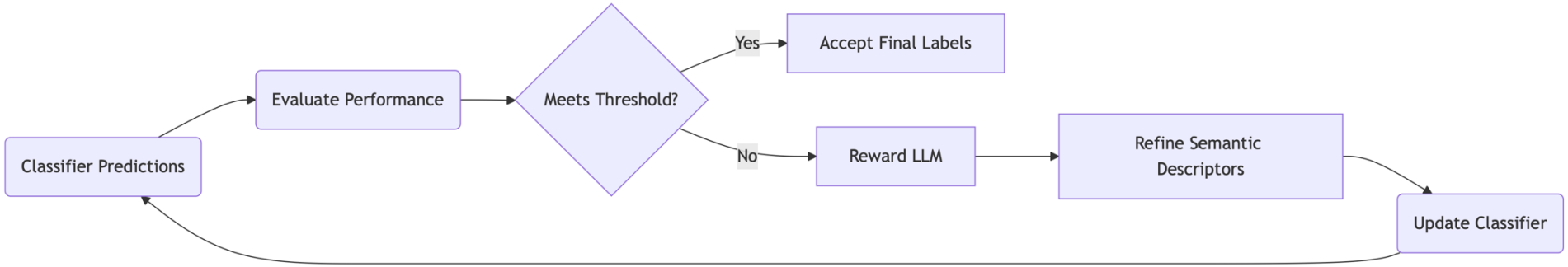
We work with existing Archetypes!

How do we know they are any good? (Based on e.g. Human Evaluation and Similiarity Matching)

⇒ ML Principle “Garbage In, Garbage Out” Applies

Solution:

- Add Feedback Loop to Improve Archetypes based on Classifier Performance



Evaluation

Datasets*

Dataset	# Classes	~ # docs
SetFit/20_newsgroups	20	18 800
sh0416/ag_news	4	127 600
fancyzhx/dbpedia_14	14	630 000
arxiv-community/arxiv_dataset	~200 + arXiv categories	1 700 000

Metrics

- **Classification:** Accuracy, Precision, Recall, F1 (macro)
- **Efficiency:** Samples/s, Latency, Throughput, Cost

* Currently used but for final evaluations datasets after LLM Knowledge Cutoffs need to be chosen

Evaluation cont‘d

Models

- **LLM’s (zero and few shot)**
  - Deep-Reasoning (GPT-4 O3, Claude 3.7 Sonnet, Gemini 2.5 Pro)
  - High-Performance (GPT-4o, Claude 3.5 Sonnet, Mixtral 8×22B (176 B))
  - Mid-size (~7 B) (Llama-3-8B (8 B), mistralai/Mistral-7B-Instruct (7.3 B))
  - Tiny (≤3 B) (google/gemma-2b (2 B), Llama-3-1B (1 B))
- **Supervised Baselines**
  - google/electra-base-discriminator (110 M)
  - roberta-base (125 M)
  - microsoft/deberta-base (139 M)

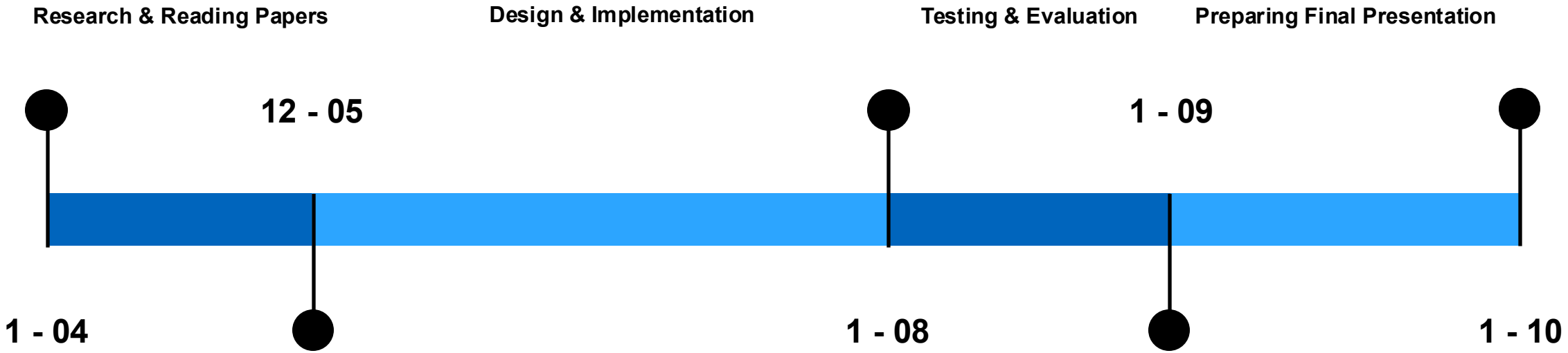
Expected Outcomes

Archetype viability: confirm that context-driven semantic archetypes can reliably label domain documents with minimal manual effort (RQ1).

Feedback benefit: show that a simple reward loop improves both classification accuracy and archetype quality (RQ2).

Competitive efficiency: demonstrate that the archetype pipeline approaches supervised and zero-shot LLM performance while using far fewer resources (RQ3).

Timeline



*The Thesis will be written over the whole 6 months

Moritz Steigerwald
Master Information Systems

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.

wwwmatthes.in.tum.de