Aalto University
MS-E2177 - Seminar on Case Studies in Operations Research

Semantic risk clustering

Interim report

Santeri Paljakka (Project Manager)

Essi Nikula

Henrik Purokoski

Jaakko Paavilainen

Antti Kärkkäinen

April 16, 2026

Contents

1 Changes in project objectives and scope

There are no major changes to the project objective or scope. After the submission of the project plan, the project team has discussed the project progress and outcomes continuously with the client, avoiding any mismatch between expectations and project team work. The objective for this project remains to be designing a semantic risk clustering pipeline for risk descriptions in risk registers, while comparing the performance of different technologies and design choices. The finished outcome contains clear presentation of the results, limitations and capabilities, and recommendations for building an usable and reliable semantic risk clustering tool.

2 Project Status

The project tasks are contained in two main categories: literature review and empirical testing. The first half of the project was spent on the literature review, to first build an comprehensive idea of the subject and current developments in the field. The methods tested in the empirical phase for preprosessing, embeddings, diamensionality reduction and clustering algorithms were then selected based on the the literature review.

While working on the literature review and empirical testing, we have updated the documentation towards the final report and presentation. Our goal is to get all necessary results and have the first version of the final report ready before the first of May. This way, we have time to complete final refinements and work on the presentation of the results in May.

2.1 Literature review

The intial literature review was finalized in the beginning of March. After that, we have been able to revise the structure and add missing pieces. The literature review now contains all parts of the semantic clustering pipeline, including data, preprocessing, embedding, similarity metrics, dimensionality reduction and clustering algorithms. In the process, the literature review has expanded, and it will be summarized and scoped for the final version.

2.2 Empirical testing

The empirical testing was started in March after the first scoping based on the literature. Table 1 contains the scoped methods for the pipeline evaluation.

Currently, the project team has set up the code base to enable easy configuration of test scenarios. The pipeline output is evaluated using numerical scores, such as the Silhuette score, Davies-Bouldin index and visualizations in two dimensions. Further development in the following weeks includes the preprosessing phase and running a grid search for all the combinations of selected parameters.

Table 1: Methods selected for pipeline evaluation

Process Step	**Tested Method**
Preprocessing	Homogenizing the input data
	No preprocessing
Embedding models	F2LLM-v2
	Text-embedding-3-large
	Harrier-oss-v1
Dimensionality Reduction	PCA
	UMAP
	No dimensionality reduction
Distance Metrics	Euclidian distance
	Cosine similarity
Clustering Algorithms	Agglomerative hierarchical
	HDBSCAN

3 Schedule

The updated schedule is presented in Figure 1. There are no major changes, but we updated the placement of the final report block to better reflect actual timing. While the literature review can be seen as a separate task to scope and research the problem, it is also an integral part of the final report. Therefore it is more accurate to say that the composing of final report started already at week 5.

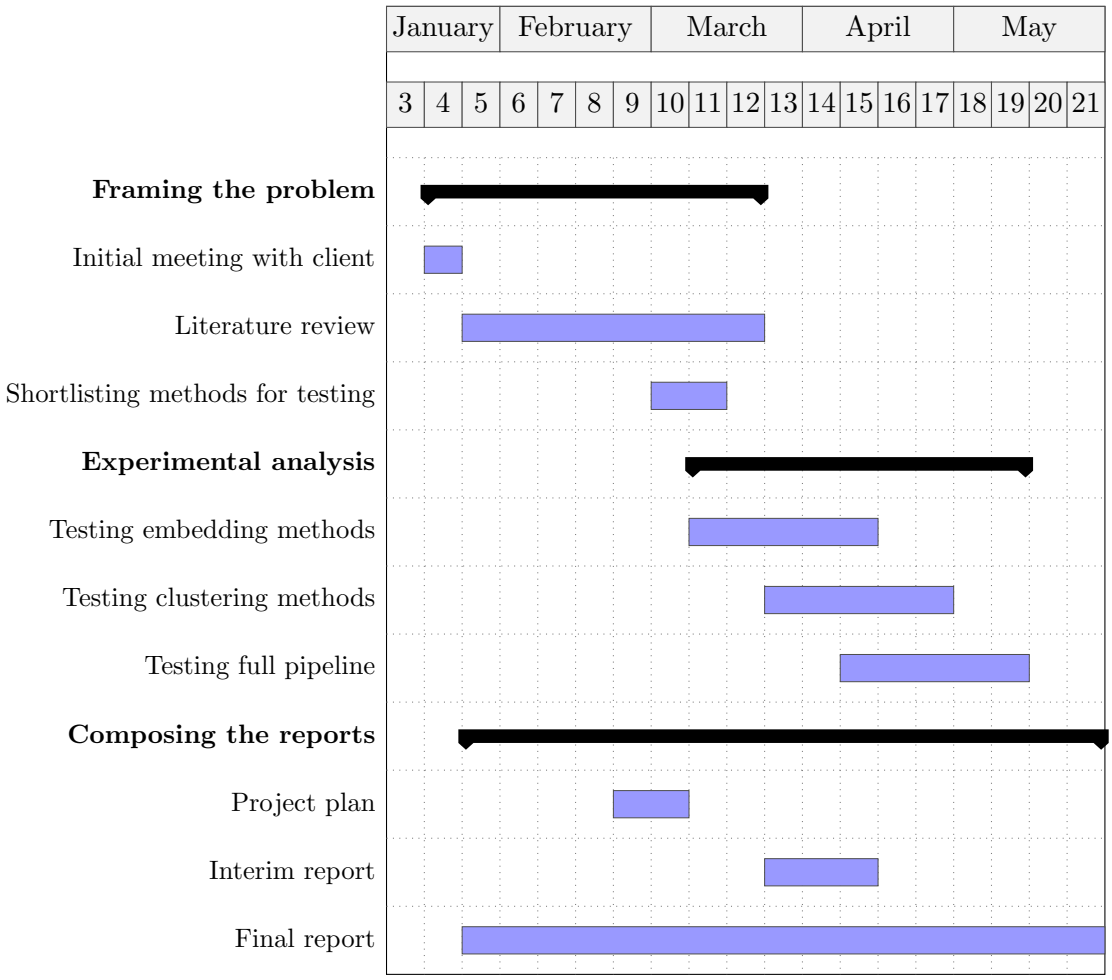

Figure 1: Project schedule

4 Risks

As a result of the constant risk monitoring, one new risk regarding parameter space was identified and added to the risk table. The team continues to monitor and identify new risks as the project proceeds.

Risk	Effect	Likelihood	Impact	Mitigation
Problem definition too broad	No effective solution for the client; workload and resources stretched	Medium	High	Clear planning and communication with the team; clear goal definition in the planning phase
Clusters not intuitive for the risk context	Found solution not useful for the client	High	Medium	Use clustering and validation methods that encourage intuition in addition to cluster correctness
Scheduling problems	Some features might not have been implemented	Medium	High	Clear schedule and task allocation; regular meetings with the team
Communication challenges with the client	Solution does not meet the client's expectations	Low	Medium	Clearly define project scope and goals with the client; regular check-ins
Communication challenges within the team	Workflow not efficient / responsibilities not properly distributed	Low	Medium	Regular intra-team meetings; encourage questions; project manager monitoring and task assignment
Insufficient risk data	Model does not properly identify real-world risk clusters	Medium	Medium	Conduct exploratory data analysis on the given data; identify possible issues
Inadequate cluster validation	Model performance cannot be validated; model not satisfactory for the client	Low	High	Broad literature review of clustering validation methods
Inadequate parameter space	Suboptimal clustering outcomes that fail to capture meaningful patterns in the data	Low	High	Broad literature review of previous research on similar pipelines

Table 2: Risk table describing the most relevant risks