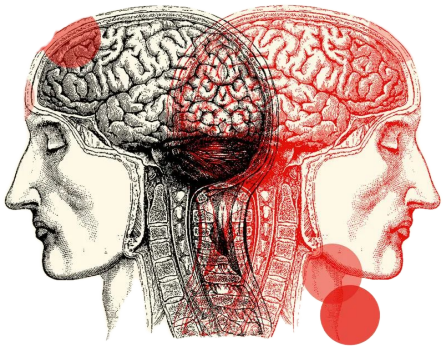
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Francisco Ribeiro | up202104797
Marisa Azevedo | up202108624
Toni Grgurevic | up202408625

Mental Disorders

Search System

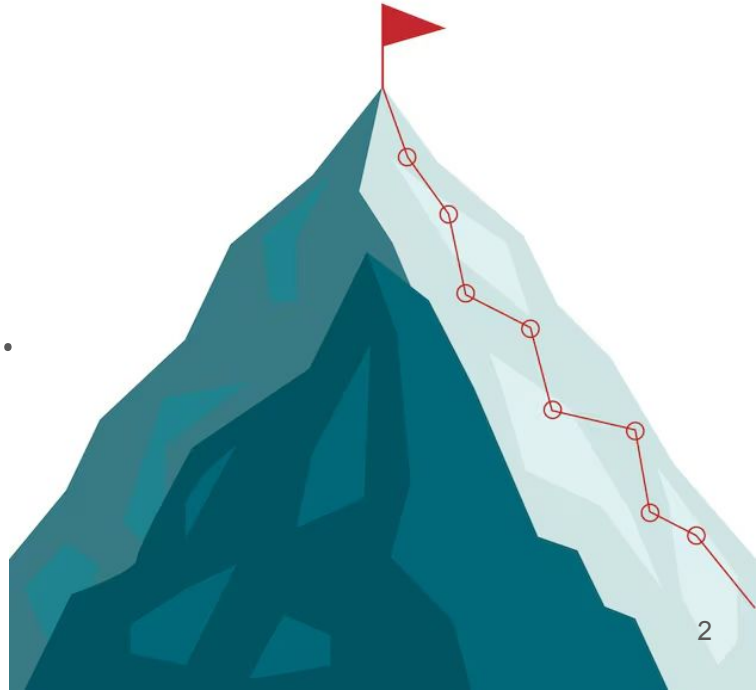


Objectives

The project consists in design and implementation of an information processing and retrieval system of Mental Diseases.

Milestone focus:

- Improve the IR system for mental health disorders.
- Explore relevance feedback.
- Using semantic search to build queries.
- Evaluate and compare queries.
- UI and GUI.

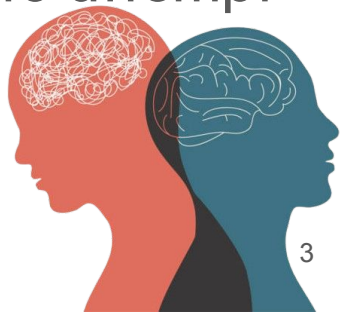


Rewind

Overview:

- Each document represents a mental health disorder;
- Includes structured (e.g. page_views) and unstructured data (e.g. symptoms, content).

The previous stage of the project introduced a basic version of the search system exploring lexical search. In this phase, we worked on enhancing the system by exploring new features in order to attempt getting better results. The main highlights are …



Semantic Search

Definition: Using dense vectors to facilitate semantic search by matching documents based on contextual meaning rather than exact keyword matches, enhancing search accuracy and relevance.

We used *DenseVectorField* type for the storage of dense vectors. We generated, using a Sentenced Transformers model, embeddings for all of our documents. For this task, we embedded all the fields that contained important text (title, content, causes, symptoms, ...) in only one field “vector”.

For the schema, besides adding this changes, we used the same filters as the final schema for M2.

Query Configuration

Simple semantic search - Consists in embed the query compare to the dense vector field, and retrieve the nearest neighbors.

```
params = {
  "q": f"{{!knn f=vector topK=25}}{embedding}",
  "fl": "name,link,score",
  "rows": 25,
  "wt": "json"
}
```

Hybrid semantic+lexical search - Use the same lexical search from M2, but, as a "*bq*" (boost query), where the KNN result is treated as an additional boost factor, enhancing the scoring of documents that are both semantically similar and relevant to the lexical query.

```
params = {
    "defType": "edismax",
    "q": f"{query}^0.3",
    "bq": f"{{!knn f=vector}}{embedding}",
    "qf": "description^3 symptoms^2 causes^2 treatment^1.7 ...",
    "pf": "description^4 symptoms^2 causes^2",
    "fl": "name,link,score",
    "rows": "25",
    "wt": "json",
    "ps": "2",
    "ps2": "1"
}
```

Query processing

User–Feedback rocchio:

-Query vector modification based on user-classified documents

-Moving query vector closer to desired information need

$$\vec{q}_1 = \vec{q}_0 + \frac{\alpha}{|R|}\sum_{d \in R}\vec{d} - \frac{\beta}{|NR|}\sum_{d \in NR}\vec{d}$$

Pseudo–Feedback rocchio:

-Automatically refines query without explicit user input

-Assumes top k retrieved documents are relevant (in our approach k=4)

Evaluation Metrics

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Goals:

- Measure system effectiveness using precision and recall.

Key Metrics:

- Precision at K (P@K): Relevance of top results.
- Average Precision (AvP): Overall precision across ranks.
- Mean Average Precision (MAP): Aggregated AvP across queries.
- Precision-Recall Curves: Stability and performance visualization.

Evaluation Results

Query 1: “Cognitive speed”

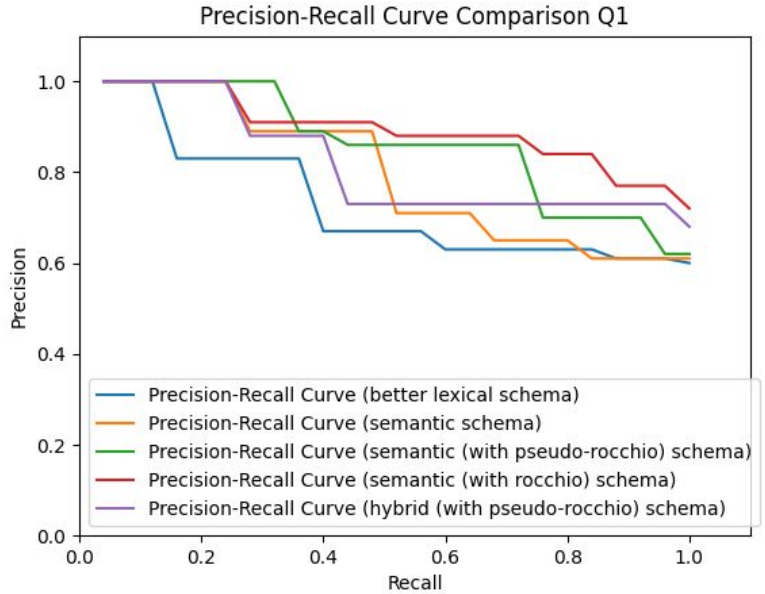


Figure 1 - *Query 1 Plot*

	Avp	P@25
better lexical schema	0.67	0.60
semantic schema	0.75	0.56
semantic (pseudo-rocchio) schema	0.81	0.60
semantic (rocchio) schema	0.86	0.72
hybrid (pseudo-rocchio) schema	0.77	0.68

Table 2 - *Query 1 Results*

Evaluation Results

Query 2: “ childhood trauma”

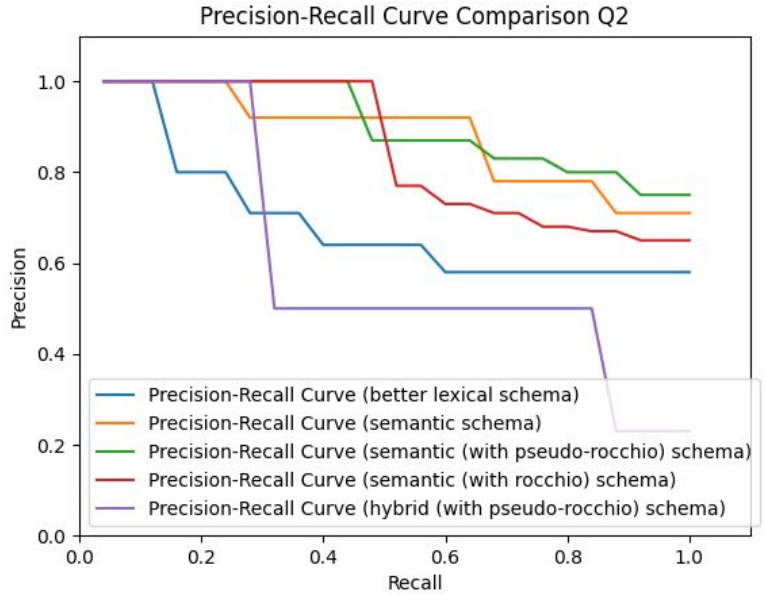


Figure 2 - *Query 2 Plot*

	Avp	P@25
better lexical schema	0.61	0.44
semantic schema	0.84	0.68
semantic (pseudo-rocchio) schema	0.88	0.72
semantic (rocchio) schema	0.81	0.60
hybrid (pseudo-rocchio) schema	0.41	0.20

Table 3 - *Query 2 Results*

Evaluation Results

Query 3: “ Improvement with behavioral therapies”

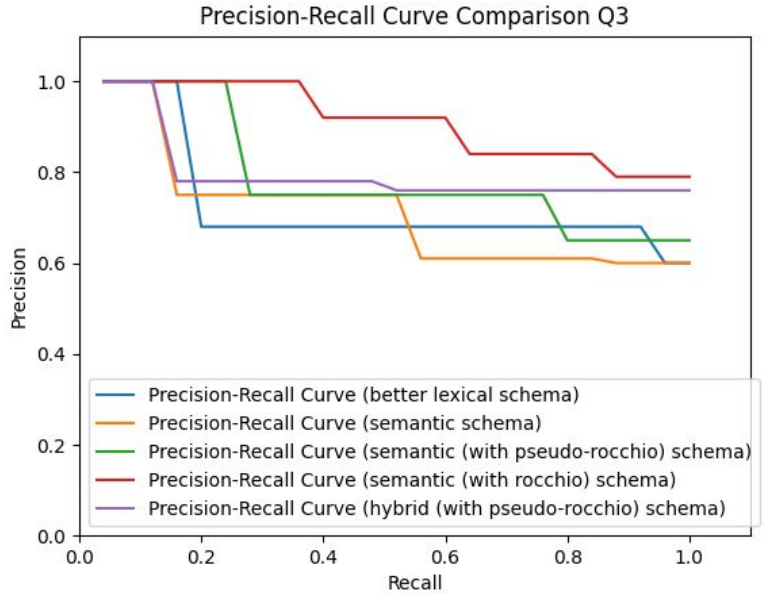


Figure 3 - *Query 3 Plot*

	Avp	P@25
better lexical schema	0.61	0.60
semantic schema	0.63	0.60
semantic (pseudo-rocchio) schema	0.69	0.52
semantic (rocchio) schema	0.89	0.76
hybrid (pseudo-rocchio) schema	0.68	0.76

Table 4 - *Query 3 Results*

Evaluation Results

Query 4: “Frequent on children”

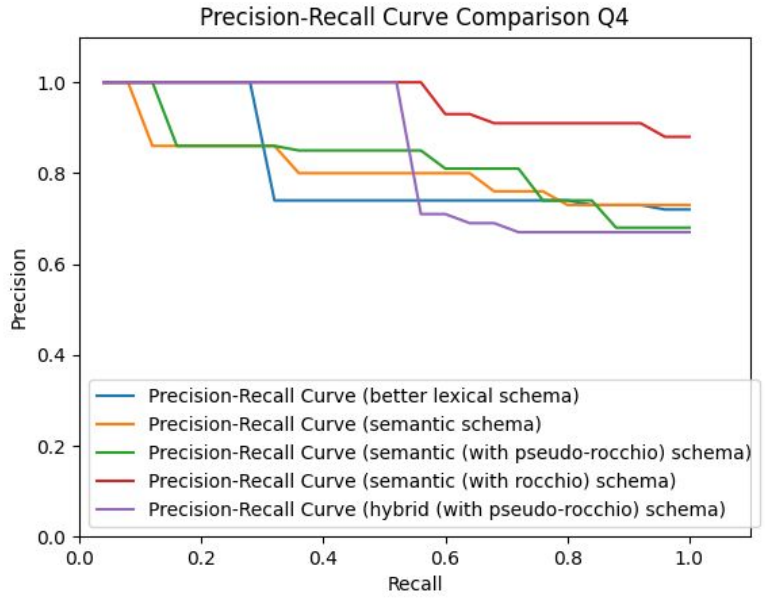


Figure 4 - *Query 4 Plot*

	Avp	P@25
better lexical schema	0.74	0.72
semantic schema	0.75	0.64
semantic (pseudo-rocchio) schema	0.78	0.68
semantic (rocchio) schema	0.95	0.84
hybrid (pseudo-rocchio) schema	0.78	0.48

Table 5 - *Query 4 Results*

Evaluation Results

Query 5: “caused by genetics inherited.”

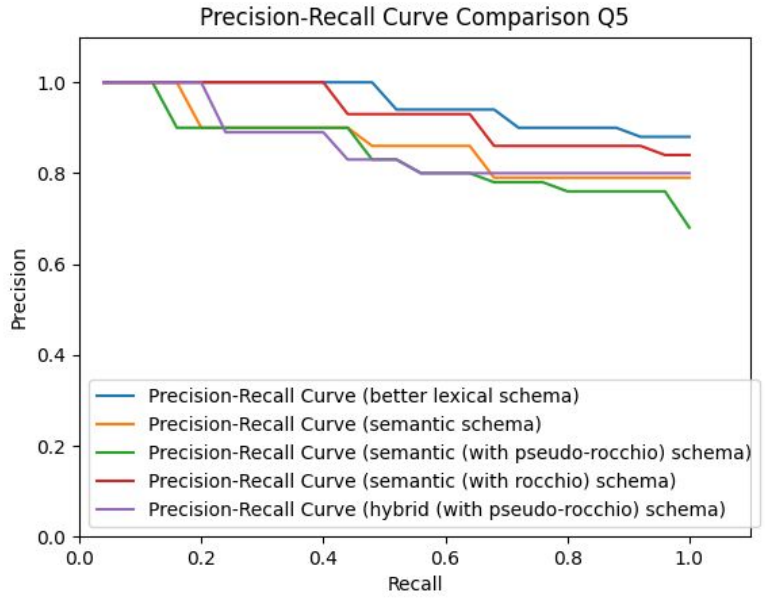


Figure 5 - *Query 5 Plot*

	Avp	P@25
better lexical schema	0.94	0.88
semantic schema	0.83	0.76
semantic (pseudo-rocchio) schema	0.80	0.68
semantic (rocchio) schema	0.91	0.84
hybrid (pseudo-rocchio) schema	0.83	0.80

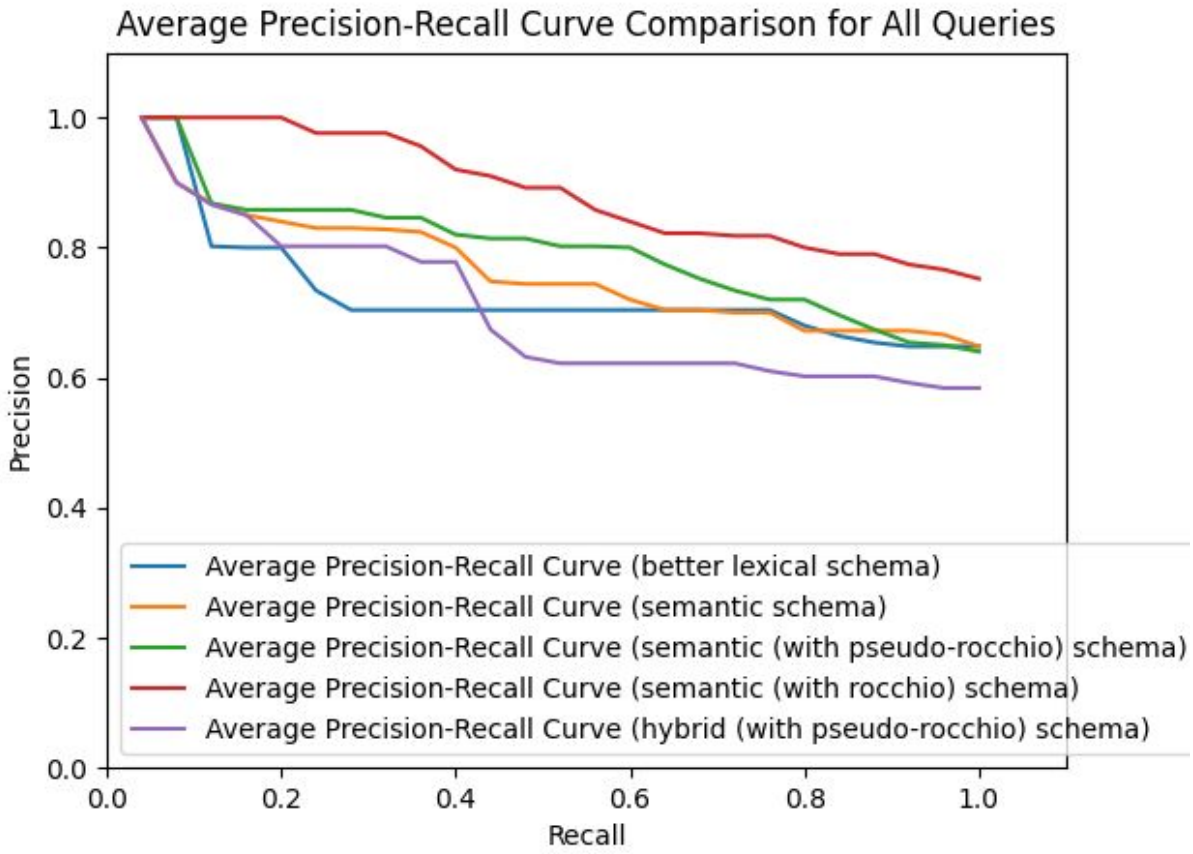
Table 6 - *Query 5 Results*

Comparative Evaluation

	Mean Average Precision (MAP)
better lexical schema	0.714
semantic schema	0.760
semantic (with pseudo-rocchio) schema	0.792
semantic (with rocchio) schema	0.884
hybrid (with pseudo-rocchio) schema	0.694

Table 7 - *MAP Scores Global*

Comparative Evaluation



Conclusion and Future Work

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Achievements:

- Successful implementation of an information retrieval system for mental health data.
- Demonstrated value of custom schema and advanced analyzers

Key Takeaways:

- Complex schema improves relevance but requires balanced optimization.

Next Steps:

- Develop a user interface for enhanced interaction.
- Improve usability and information retrieval quality for mental health data.

The end of the Powerpoint

Thanks for the attention. Do you have any question?