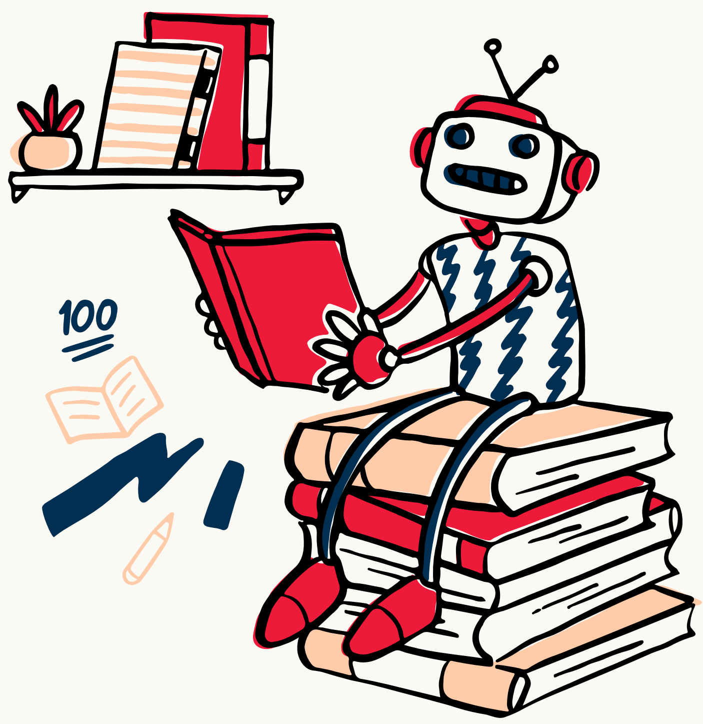
APPLIED DATA SCIENCE PROJECT

Patient Preference Studies
Classification Project

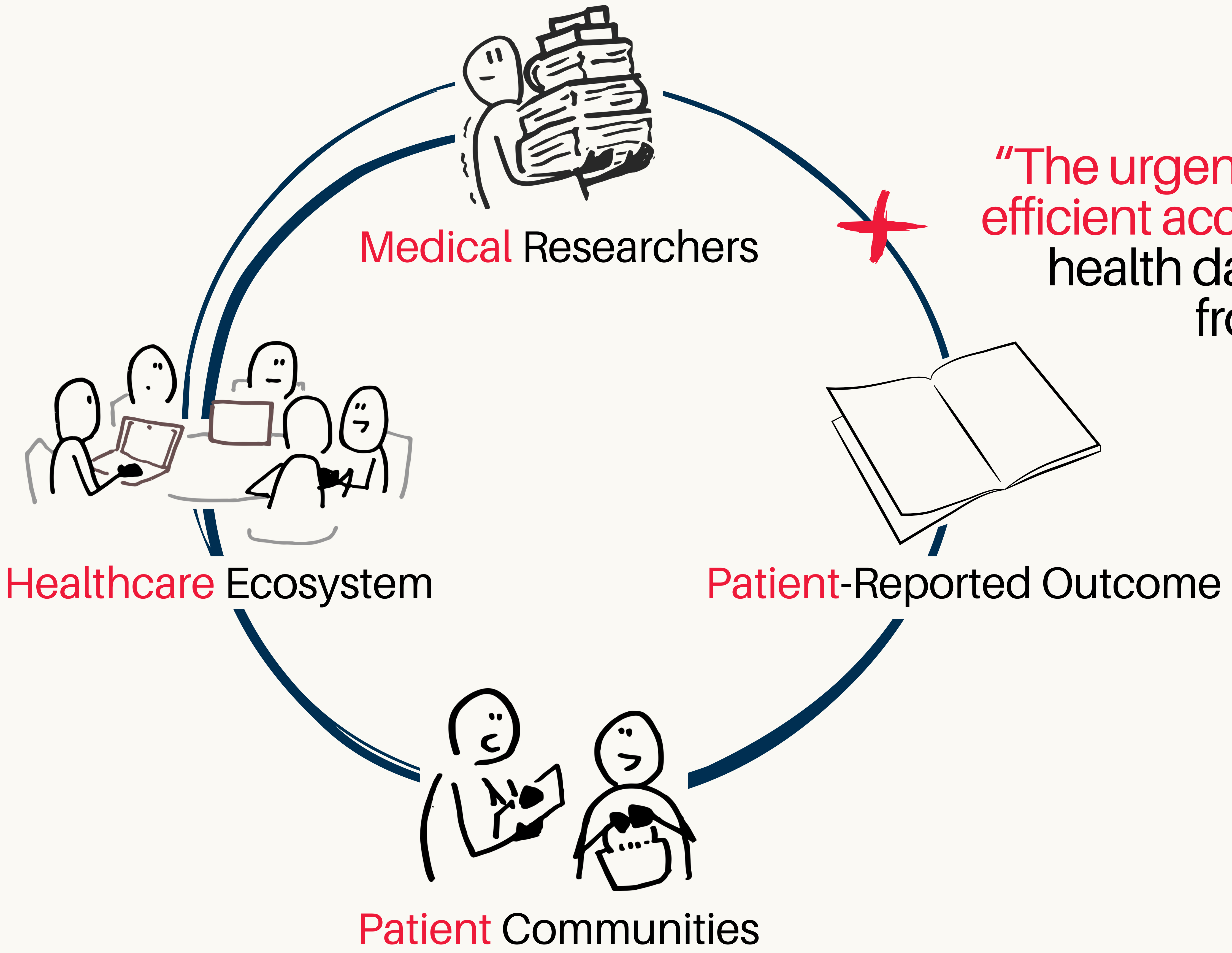


UNIVERSITÀ DI TORINO

Cesar Augusto Seminario Yrigoyen
Francesco Giuseppe Gillio

Politecnico di Torino

The **Picture**



"The urgent demand for tools that support efficient access, integration, and analysis of health data to derive actionable insights from **patient-reported outcome** and real-world evidence"

- European Commission

The **Project Value** Proposition

For medical researchers who struggle to access true Patient Preference Studies (PPS) in PubMed's vast, noisy search results—where rule-based systems fall short—our AI-driven software harnesses LLM's semantic analysis to detect true PPS and uncover relevant clinical sub-areas

UNIVERSITÀ DI TORINO

Politecnico di Torino

The **Project Value** Proposition

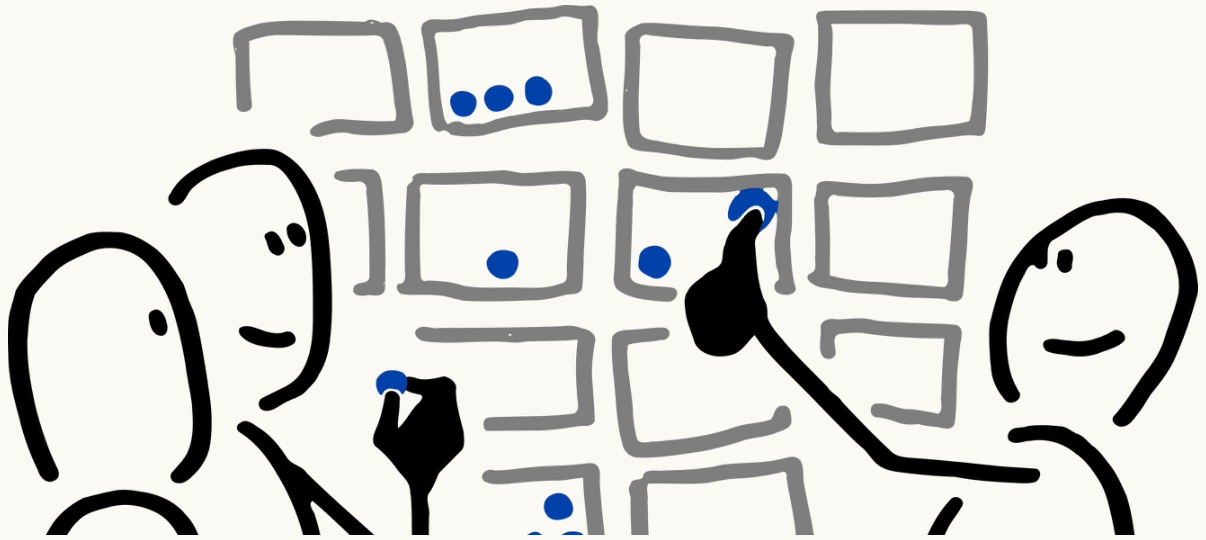
Our project supports **evidence-based healthcare** and enhances **research access** and **knowledge discovery**

3 GOOD HEALTH AND WELL-BEING

United Nations
Sustainable Development Goals
2030

4 QUALITY EDUCATION

The **Objective**



Support efficient access, collection, and validation of Patient PreferenceStudies

Develop an automated system for effective classification of PPS into relevant categories

An **easy-to-use AI software** to detect true Patient Preference Studies and uncover relevant clinical sub-areas from PubMed search results

UNIVERSITÀ DI TORINO

Politecnico di Torino

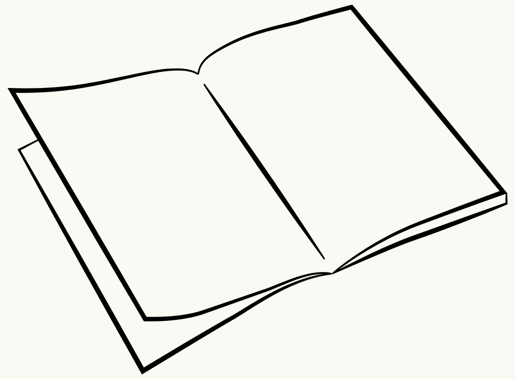
The **Data**

Data **Collectors**

Medical **Researchers**

UNIVERSITÀ DI TORINO

Data **Source**



PubMed
links queries to records via
rule-based systems (MeSH)
and lightweight machine learning
algorithms (Best Match)

Textual Data

The head-to-head comparison of diabetic patient preferences for glucose-monitoring devices.

Abstract

A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic **patient preferences** for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs. Insights from this evaluation inform the selection of preference-elicitation techniques in patient-centered healthcare research.

Politecnico di Torino

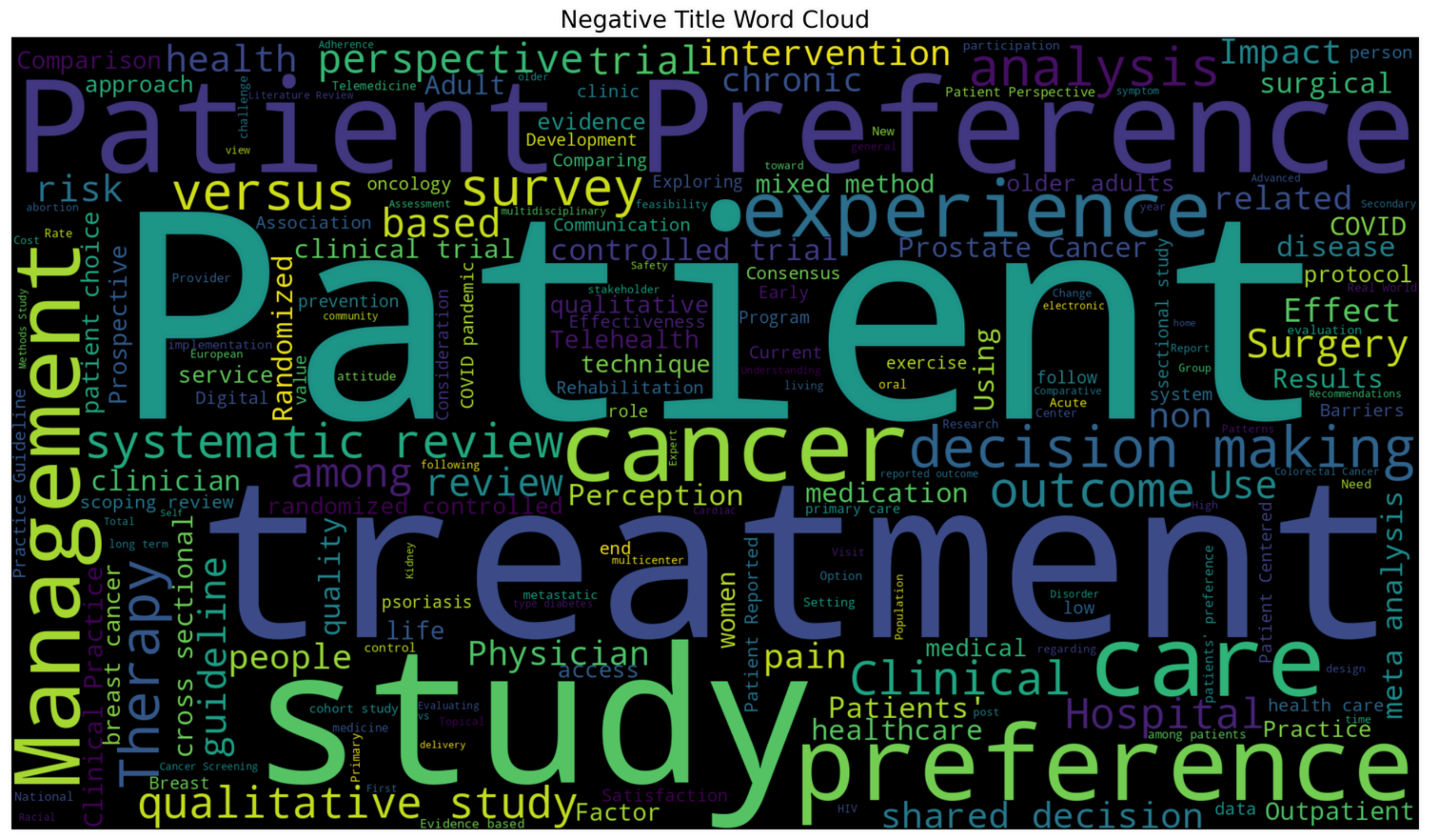
The **Data**

The head-to-head comparison of diabetic **patient preferences** for glucose-monitoring devices.

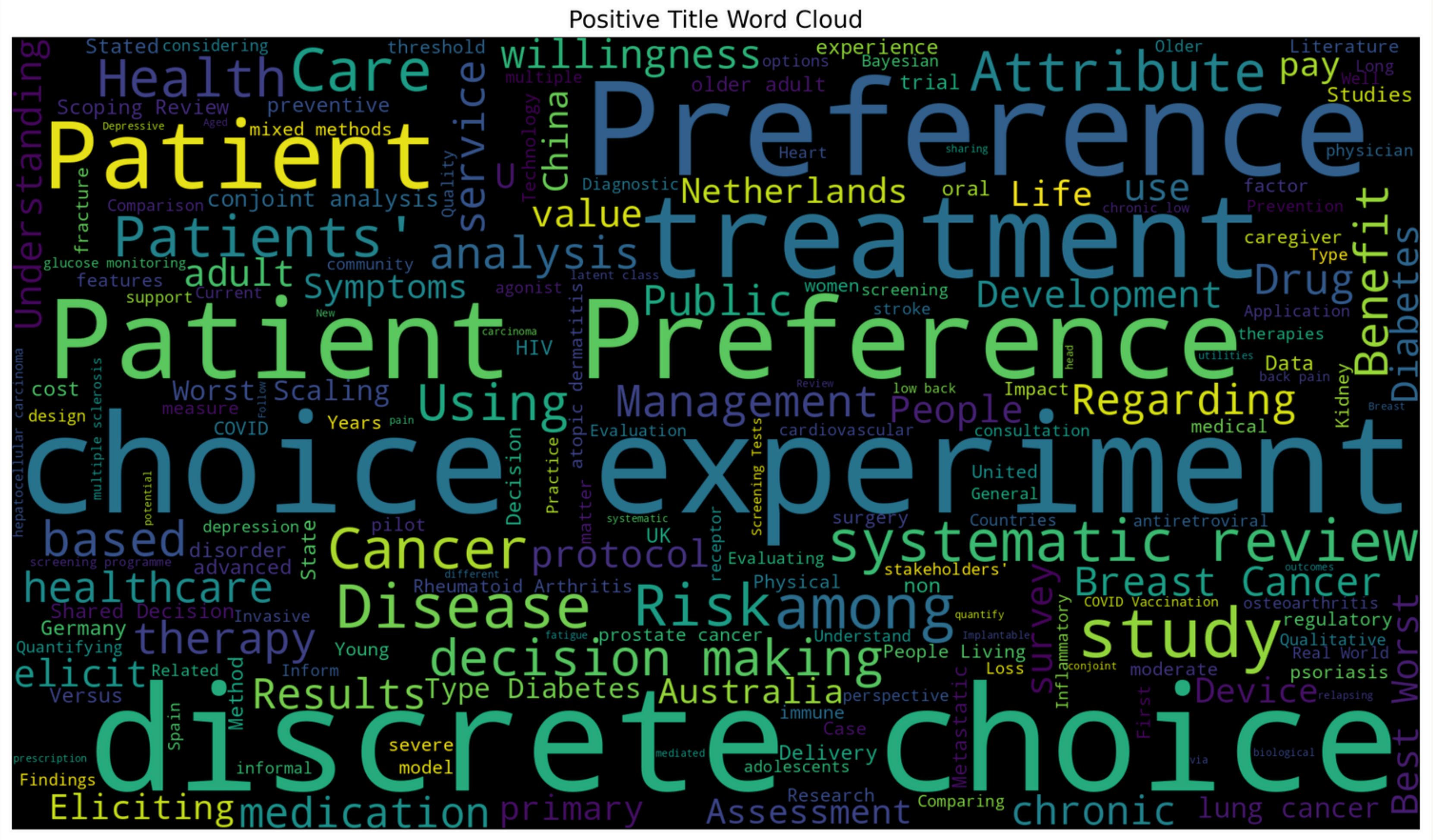
Abstract

A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic **patient preferences** for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs. Insights from this evaluation inform the selection of preference-elicitation techniques in patient-centered healthcare research.

Non Patient Preference Study



Patient Preference Study

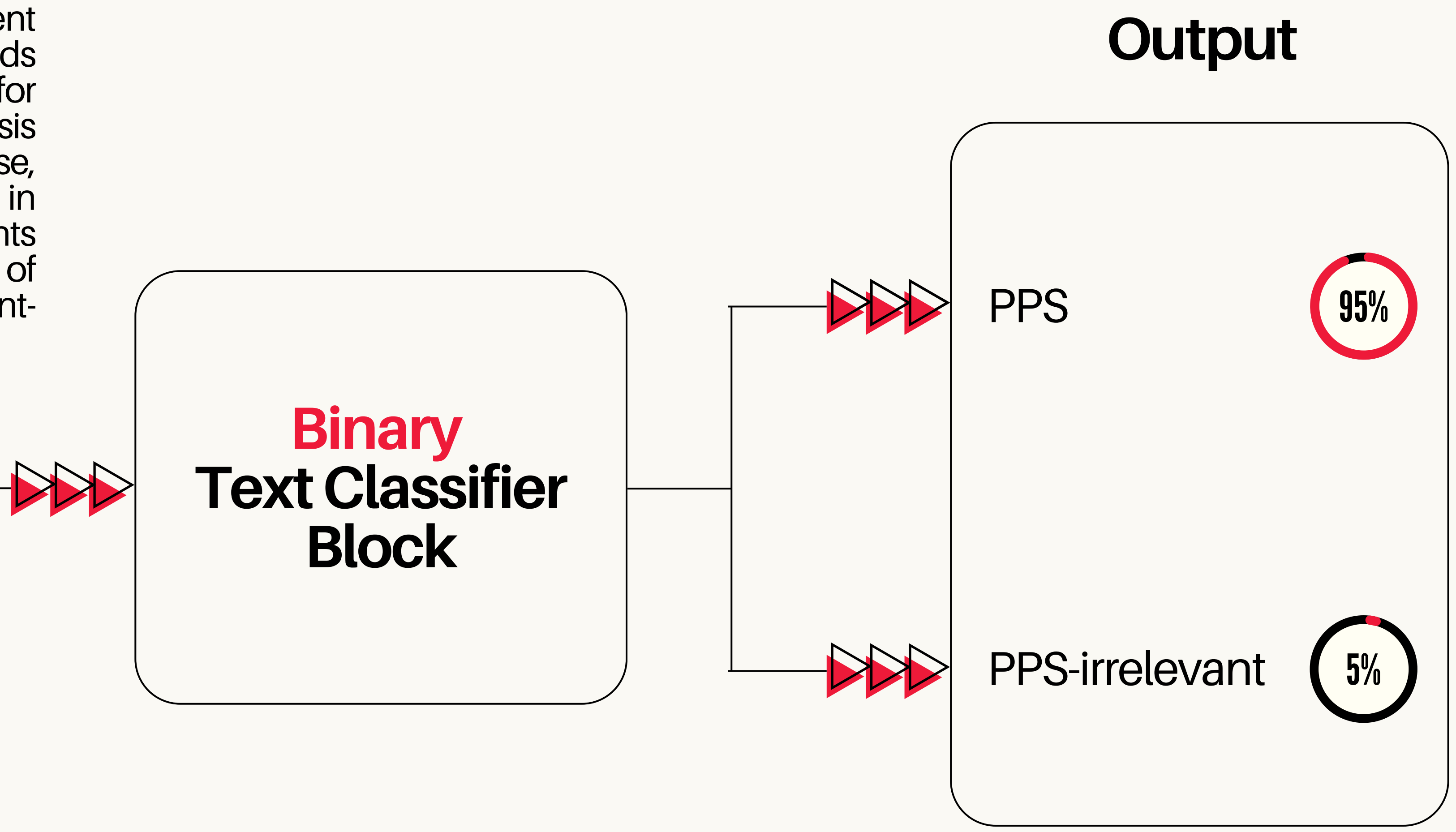


Binary Text Classification

The head-to-head comparison of diabetic patient preferences for glucose-monitoring devices.

Abstract

A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic **patient preferences** for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs. Insights from this evaluation inform the selection of preference-elicitation techniques in patient-centered healthcare research.



The **Hypotheses** and Research Questions

Large Language Models (LLMs) understand contextual relationships in text, enough to fool a human

How to adapt a Large Language Model for text classification by relevance to **Patient Preference Studies**?

Traditional ML Models (MLMs) continue to reign supreme in clinical prediction systems

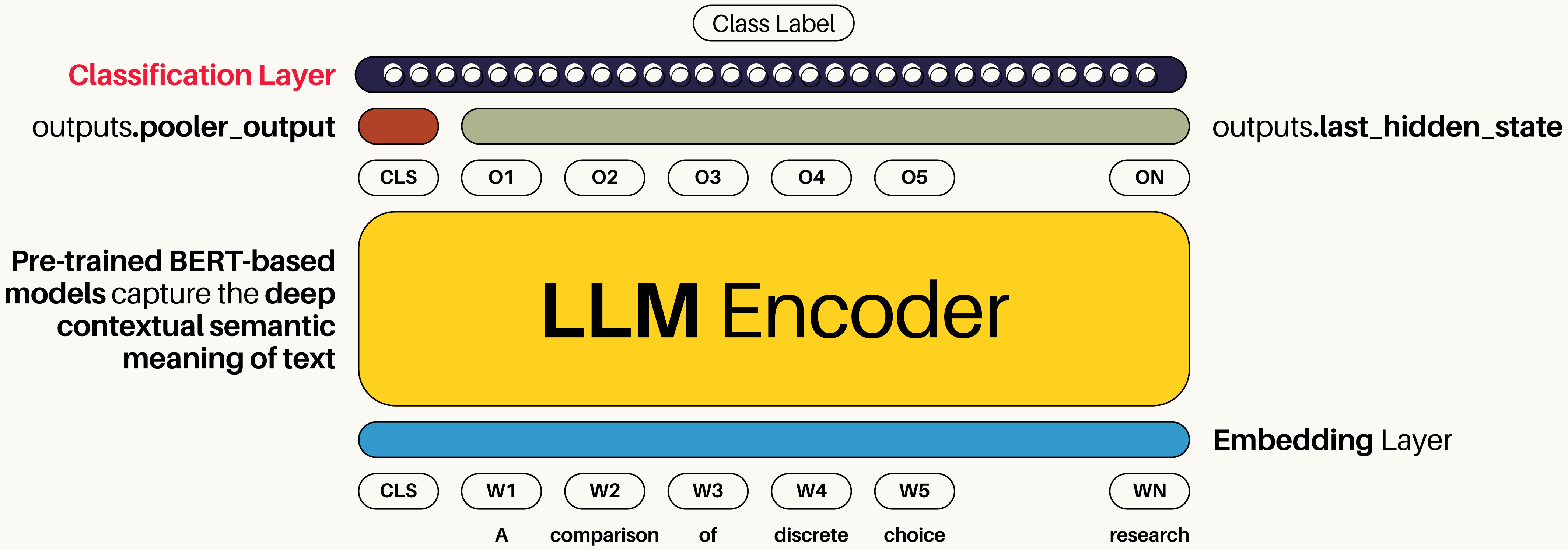
How to bring these two different worlds together into a single and innovative system?

Classification Layer on top

and hyper-parameters tuning to recognize the Patient Preference Studies pattern.

So train and test,

why not?

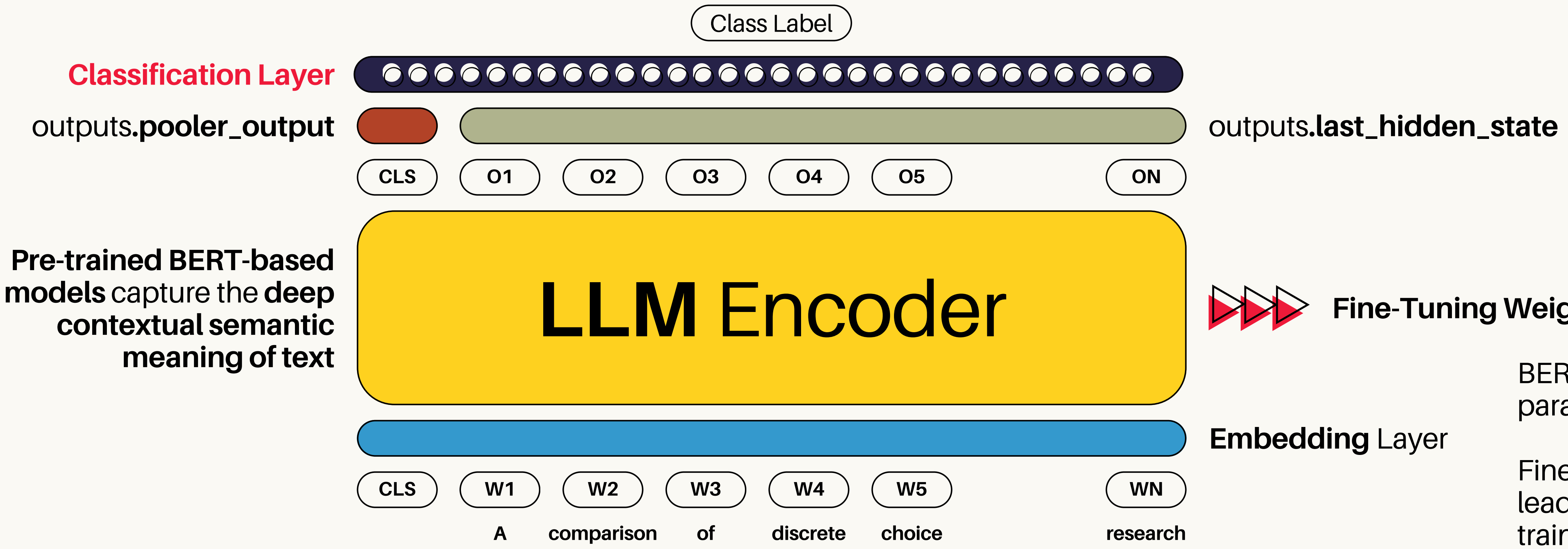


A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic **patient preferences** for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs. Insights from this evaluation inform the selection of preference-elicitation techniques in patient-centered healthcare research.

LLM for Text Classification

Dataset Size: 1.200 Papers

with fewer than 200 Patient Preference Studies,
not enough data to train a LLM from scratch,
not enough to fine-tune a pre-trained LLM
without **overfitting**



A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic **patient preferences** for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs. Insights from this evaluation inform the selection of preference-elicitation techniques in patient-centered healthcare research.

Fine-Tuning Weights or **Not**?

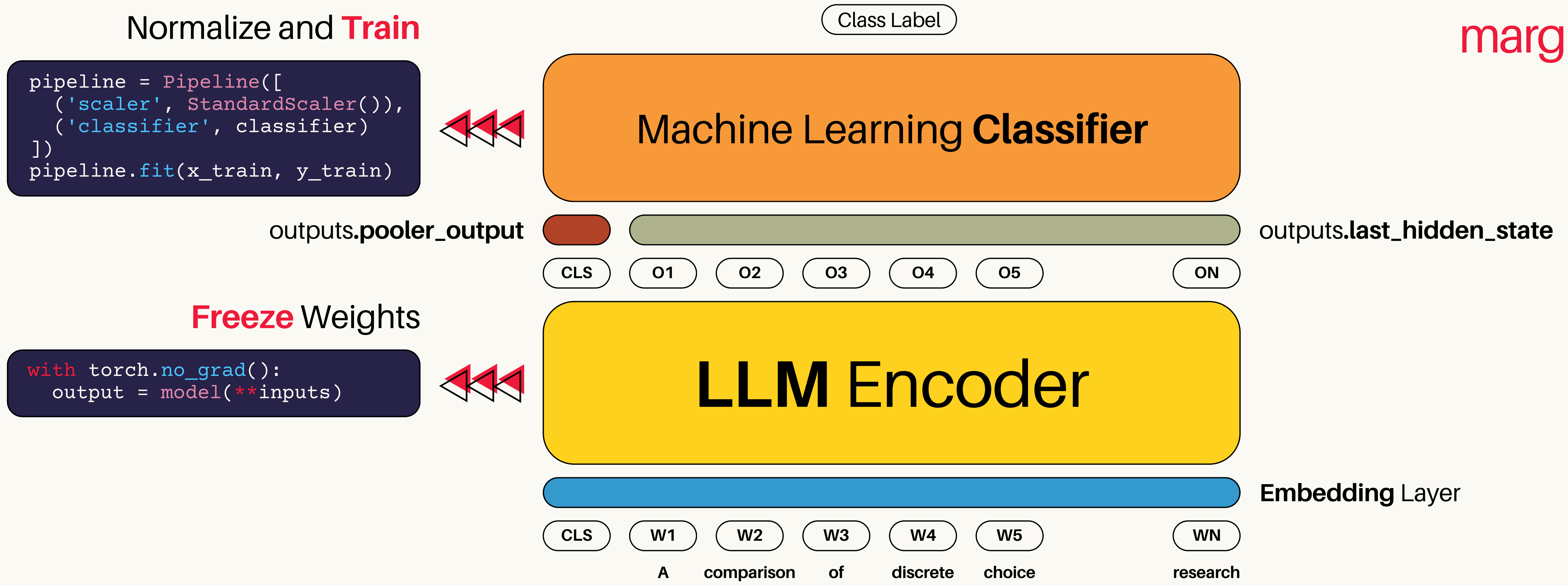
BERT hides a large number of parameters

Fine-tuning weights on a small dataset leads the model to memorize the training data (**overfitting**)

The **Hybrid** Solution

BERT-Based Model with **frozen weights** to encode medical texts into machine-readable format

Train a **Machine Learning Classifier** on top to draw and maximize the margin between classes



A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic **patient preferences** for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs. Insights from this evaluation inform the selection of preference-elicitation techniques in patient-centered healthcare research.

What to **evaluate?**

The worst outcome:
researchers lose papers on PPS

True Positive Rate (TPR)

$$\frac{TP}{TP+FN}$$ for Class 1 (PPS)

The probability that an actual positive tests positive

The target outcome:
researchers retrieve papers on PPS alone

▶▶▶ remove noise

Positive Predictive Value (PPV)

$$\frac{TP}{TP + FP}$$ for Class 1 (PPS)

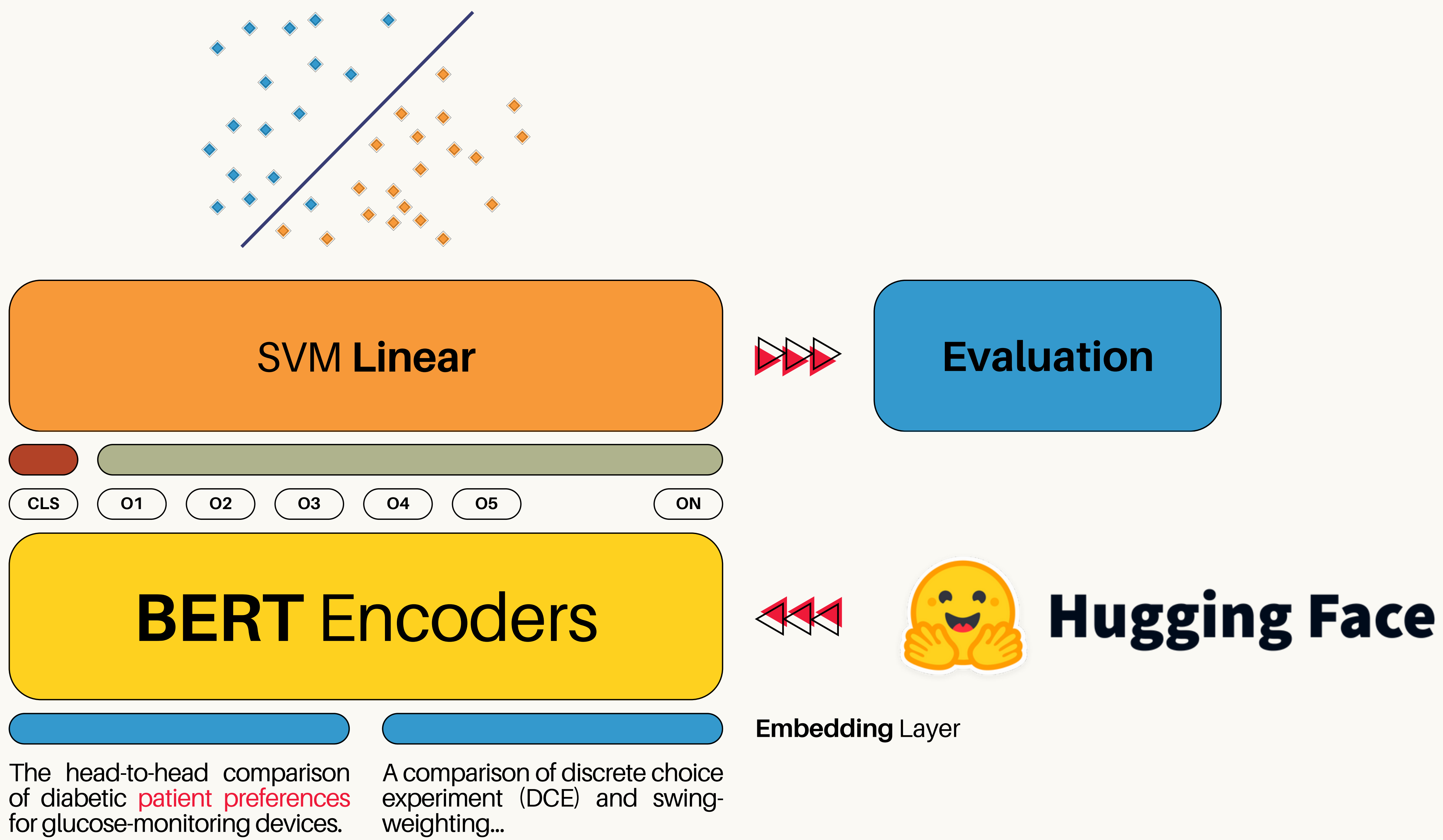
The probability that a positive test matches an actual positive

The **Experimental** Setup

The head-to-head comparison of diabetic patient preferences for glucose-monitoring devices.

Abstract

A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic **patient preferences** for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs. Insights from this evaluation inform the selection of preference-elicitation techniques in patient-centered healthcare research.

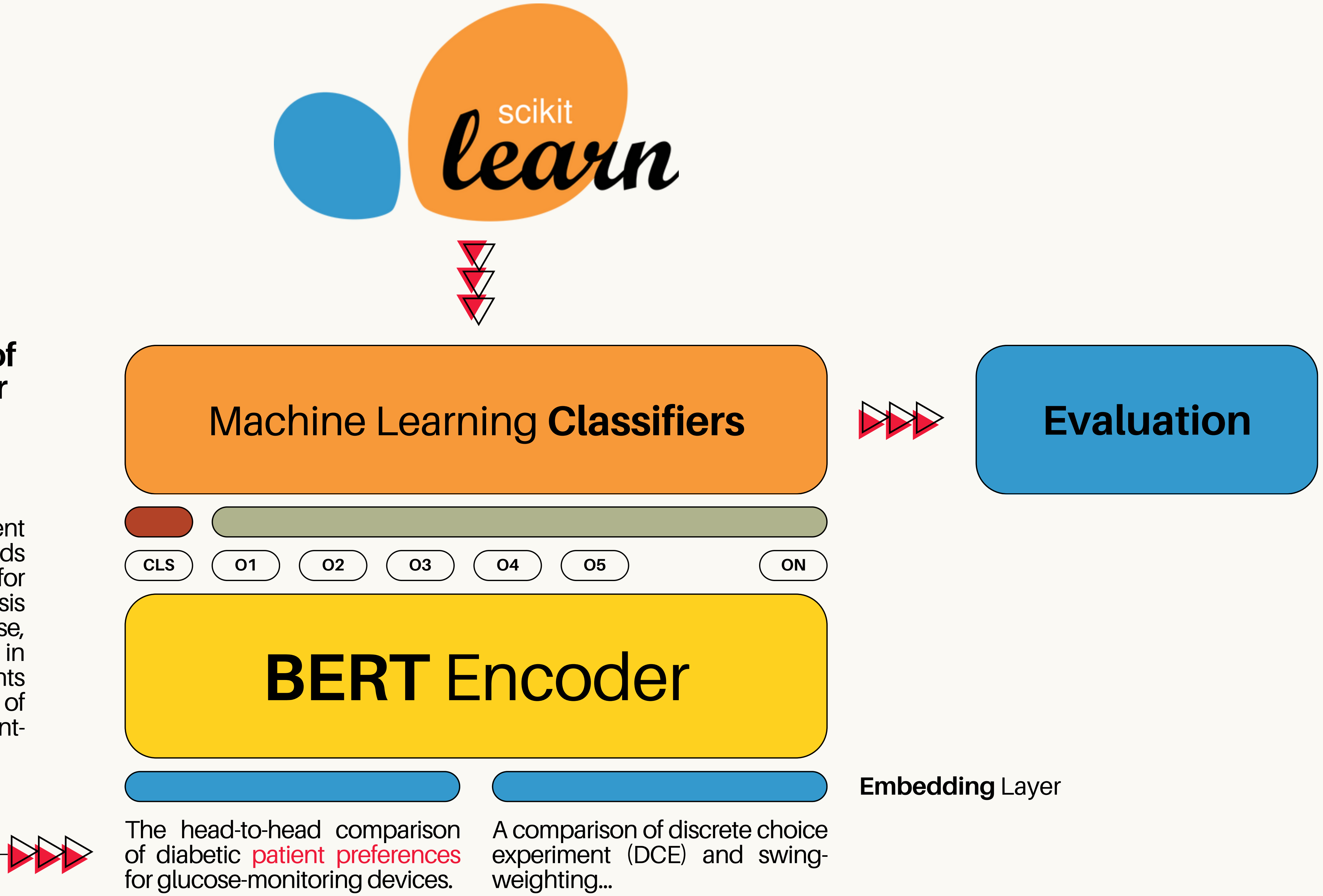


The **Experimental** Setup

The head-to-head comparison of diabetic patient preferences for glucose-monitoring devices.

Abstract

A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic **patient preferences** for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs. Insights from this evaluation inform the selection of preference-elicitation techniques in patient-centered healthcare research.



The Top Three

Hugging Face

BERT Models

BERT-Base Model	F2-PPS	F1-PPS	TPR	PPV
pubmedbert-base-embeddings	**0.833**	0.836	0.831	0.844
BiomedNLP-BiomedBERT-base-uncased-abstract	**0.821**	0.821	0.822	0.825
S-PubMedBert-MS-MARCO	0.816	0.81	0.819	0.803

scikit learn

pubmedbert-base-embeddings

Classifier Model	F2-PPS	F1-PPS	TPR	PPV
k-Nearest Neighbors	**0.895**	0.82	**0.953**	0.719
Deep Neural Network	0.893	0.892	0.894	0.882
Logistic Regression	0.891	0.891	0.889	0.884

BiomedNLP-BiomedBERT-base-uncased-abstract

Classifier Model	F2-PPS	F1-PPS	TPR	PPV
SVM (RBF Kernel)	**0.906**	0.91	0.903	**0.918**
Neural Network	0.9	0.906	0.896	0.914
Logistic Regression	0.899	0.9	0.898	0.899

The **Top** Parameters

k-Nearest Neighbours

Parameters	**Values**
K (**neighbours**)	3, 5, 8, 13, 21
Metric	euclidean, manhattan, minkowski

for **pubmedbert-base-embeddings**

High Recall Model

K-Nearest Neighbors Parameters	**F2-PPS**	**F1-PPS**	**TPR**	**PPV**
K = 5, **Metric** = euclidean	0.897	0.823	**0.953**	0.724

Support Vector Machine

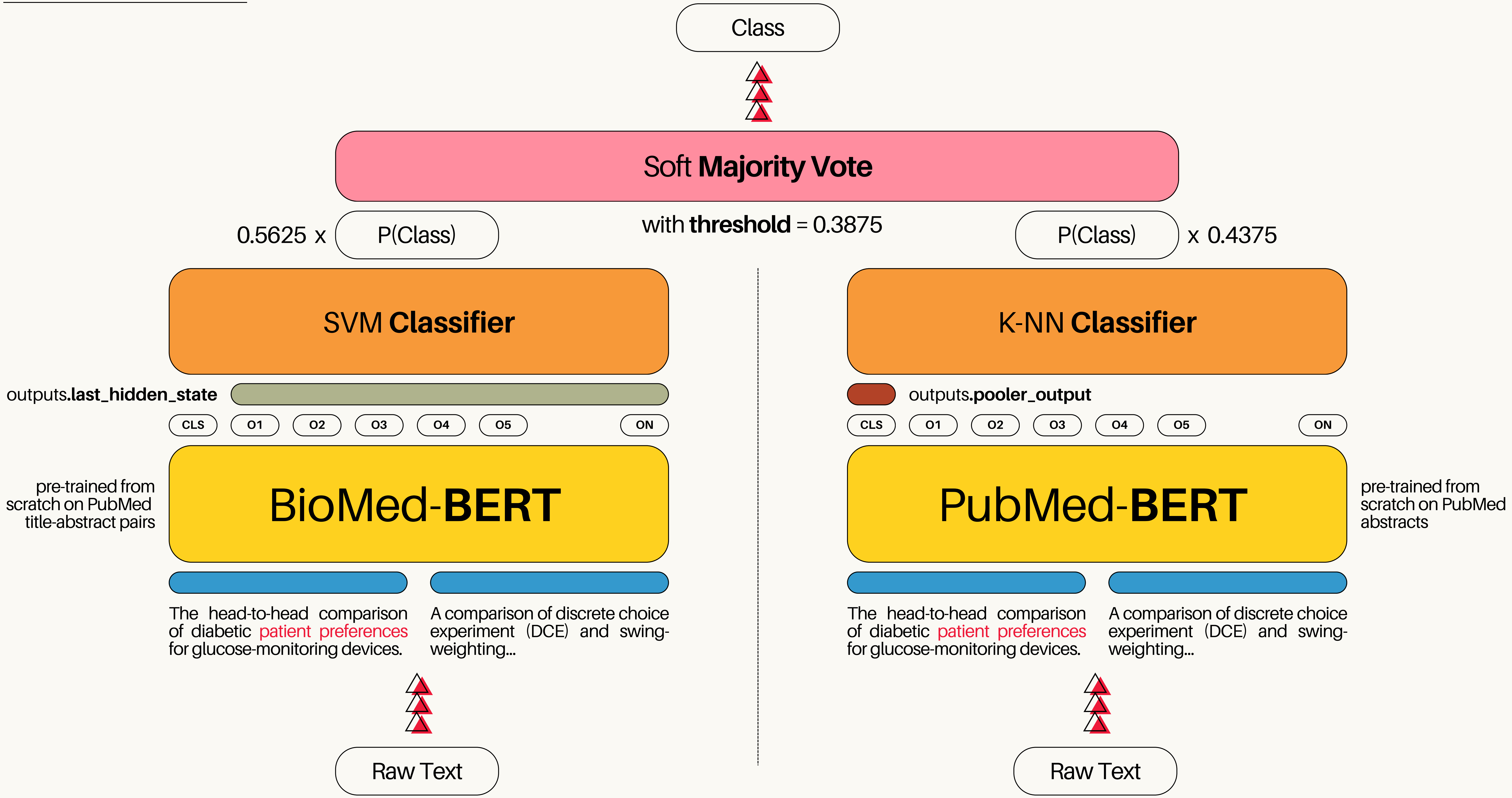
Parameters	**Values**
C (**penalty parameter**)	1e0, 1e1, 1e2
Gamma (**kernel coefficient**)	1e-5, 1e-4, 1e-3, auto, scale

for **BiomedNLP-BiomedBERT-base-uncased-abstract**

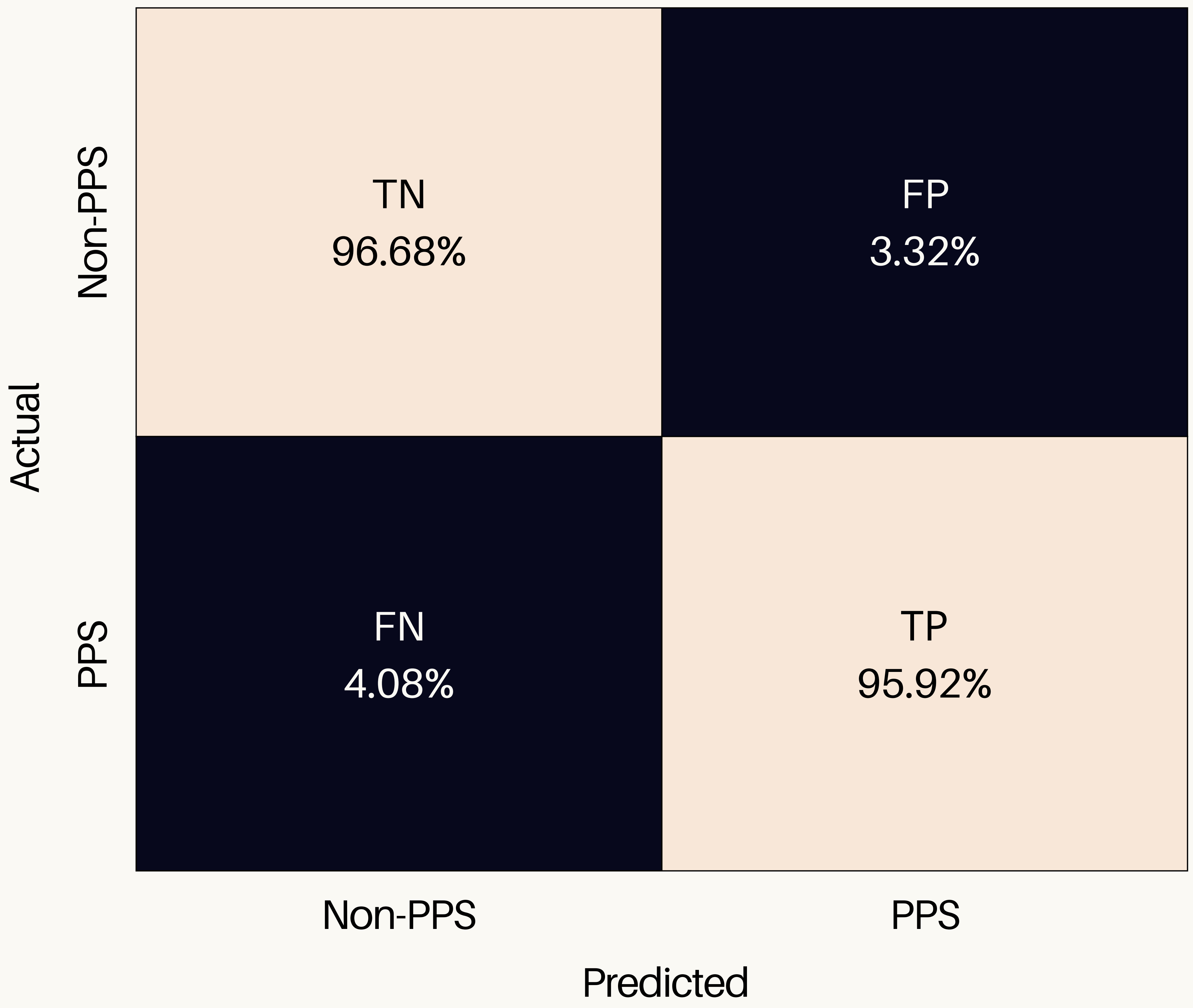
High Precision Model

Support Vector Parameters	**F2-PPS**	**F1-PPS**	**TPR**	**PPV**
C = 1e1, **Gamma** = auto	0.901	0.921	0.888	**0.957**

The **Deliverable** Model



The **Result**



The model loses 5 PPS out of 100 papers, but **removes more than 95% of the noise**

What's Next?

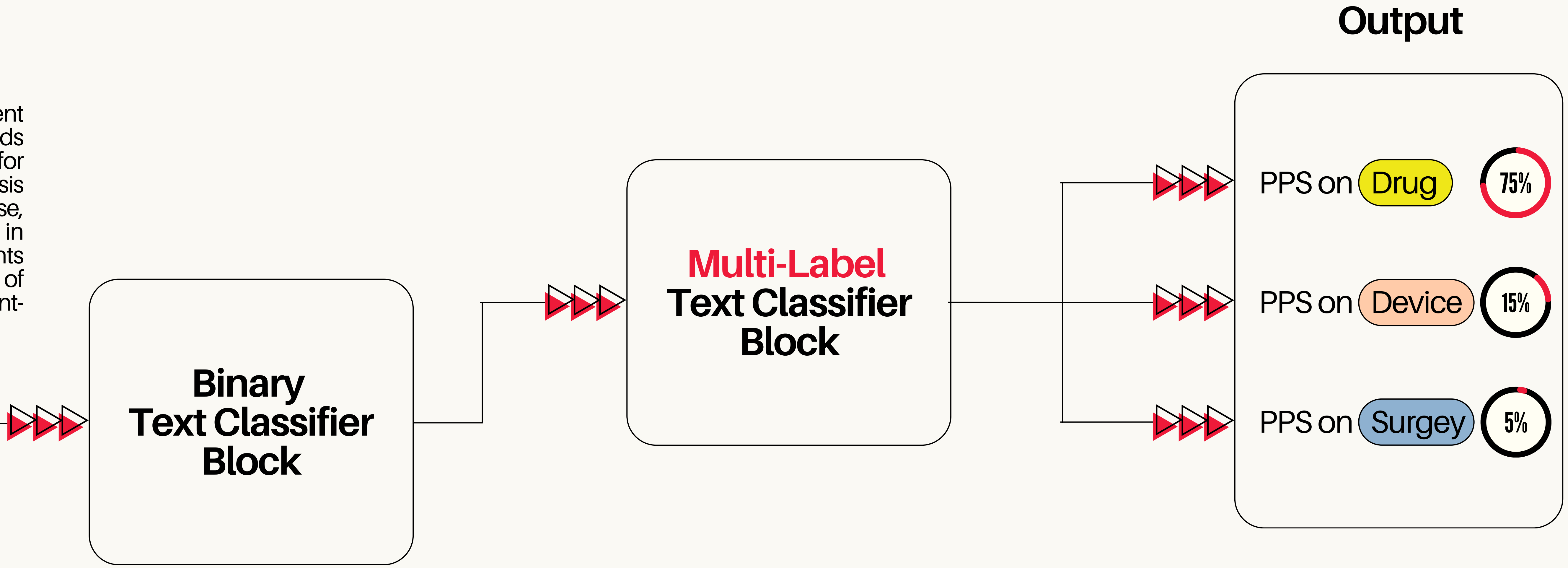
The **Multi-Label** Text Classifier

Multi-Label Text Classification

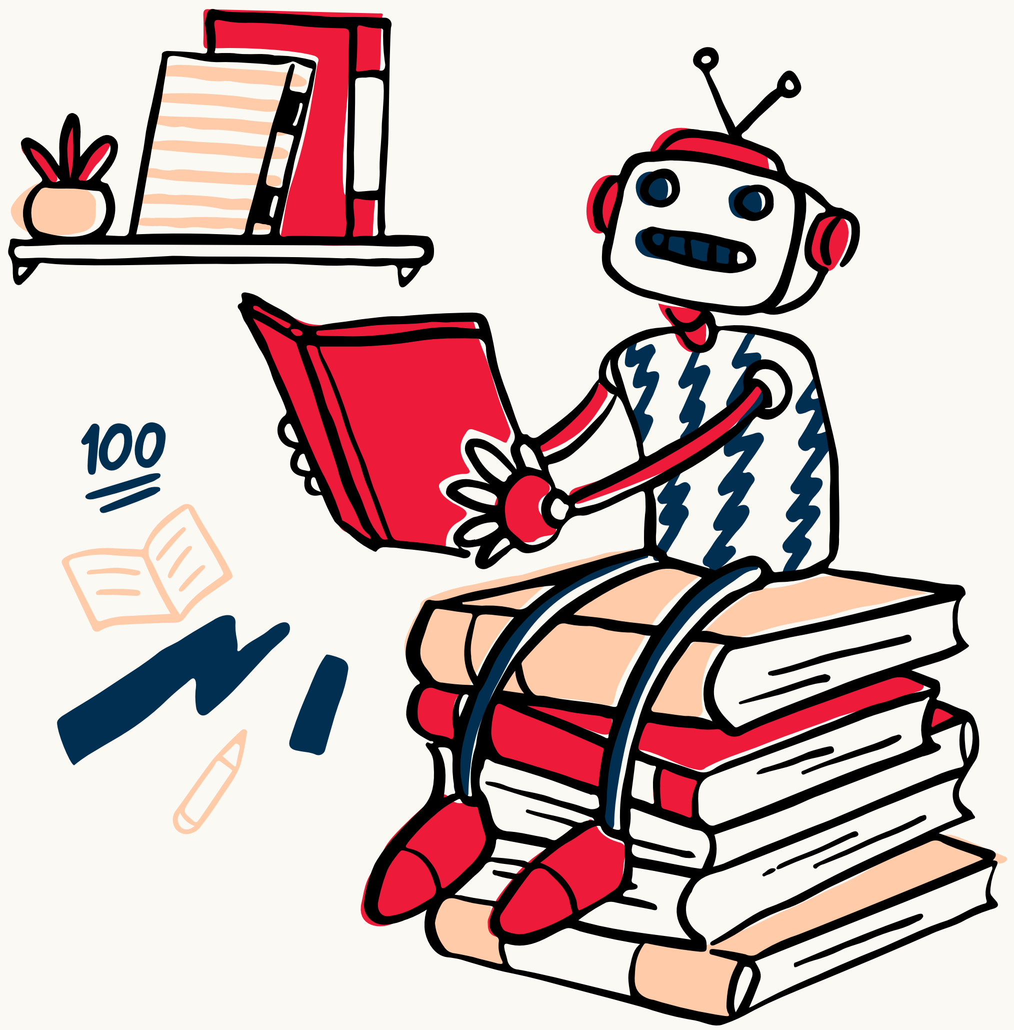
The head-to-head comparison of diabetic patient preferences for glucose-monitoring devices.

Abstract

A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic **patient preferences** for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs. Insights from this evaluation inform the selection of preference-elicitation techniques in patient-centered healthcare research.



The **Multi-Label** Classifier Methods



1. **ML Dataset and Imbalance Analysis**

- Mean imbalance ratio
- Coefficient of variation of IR
- Scumble index

2. **Data-driven approach for model selection: comparison of most used MLC**

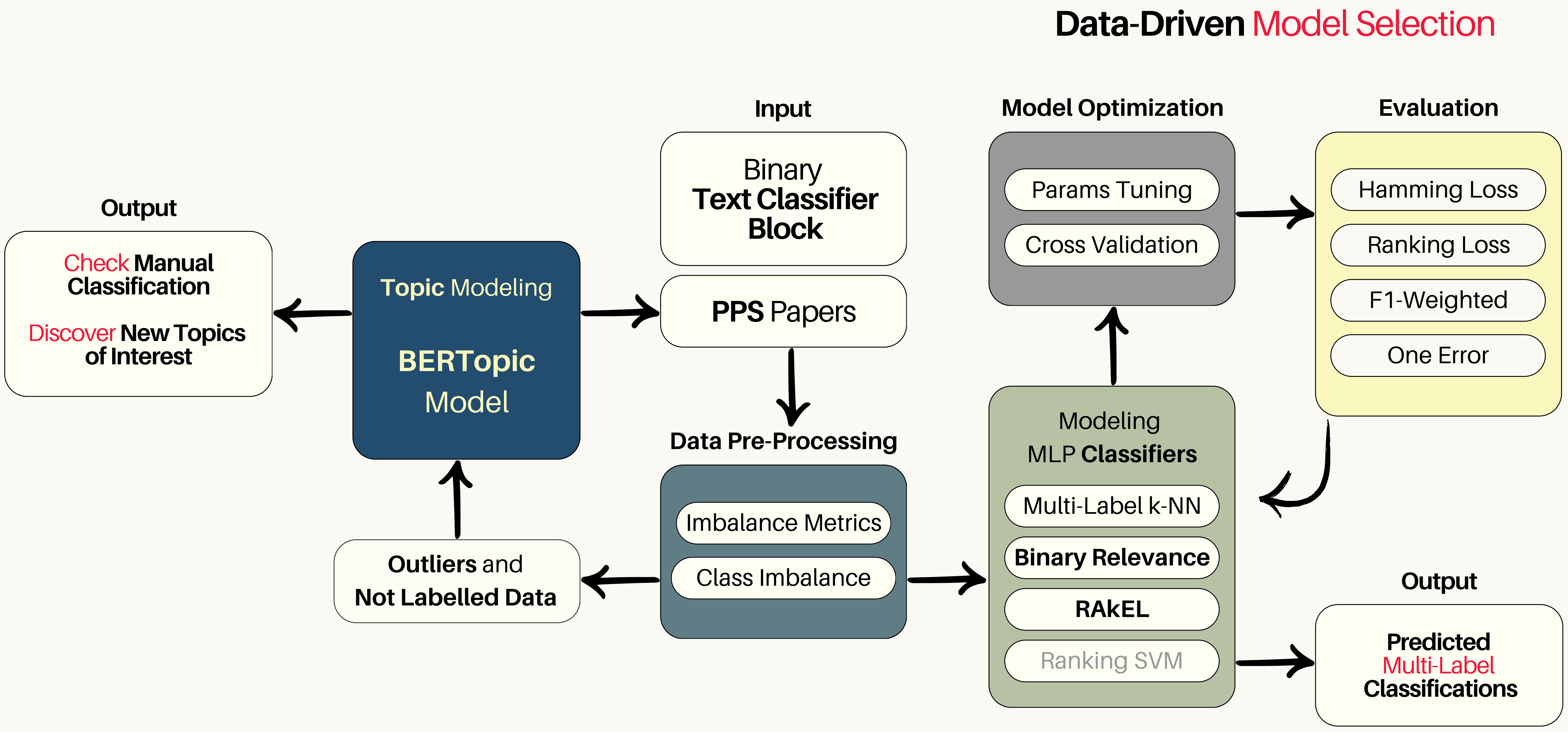
- Problem transformation (Binary relevance, RAkEL)
- Algorithm adaptation (multilabel kNN, VW ML kNN, ranking SVM)
- Ensemble models
- Cost sensitive (RAkEL)

3. **Metrics for model comparison**

- Hamming loss
- Ranking loss
- F1-score (micro, weighted)
- Coverage error
- One error

4. **Best model selection**

The Experimental Setup



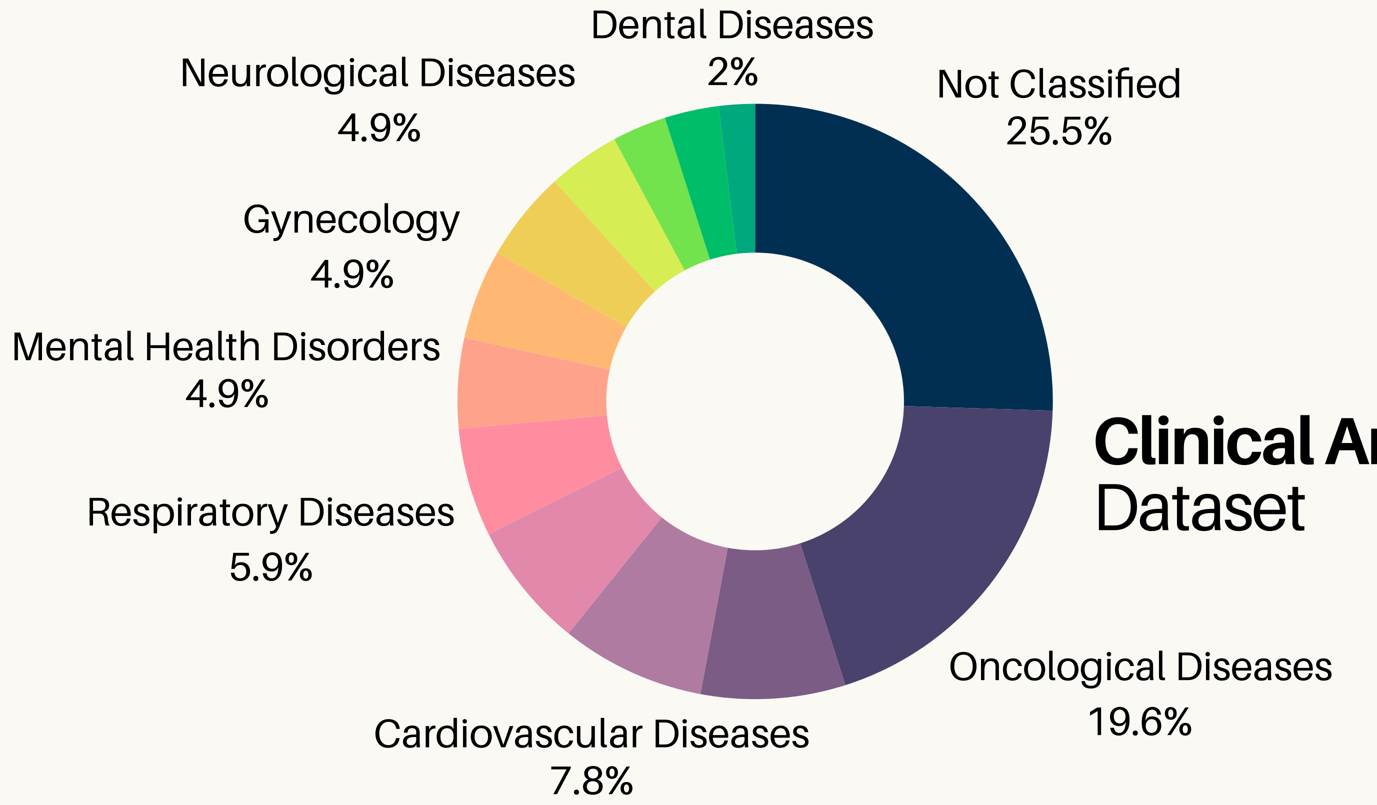
Dataset **Imbalance** Metrics

Dataset Imbalance

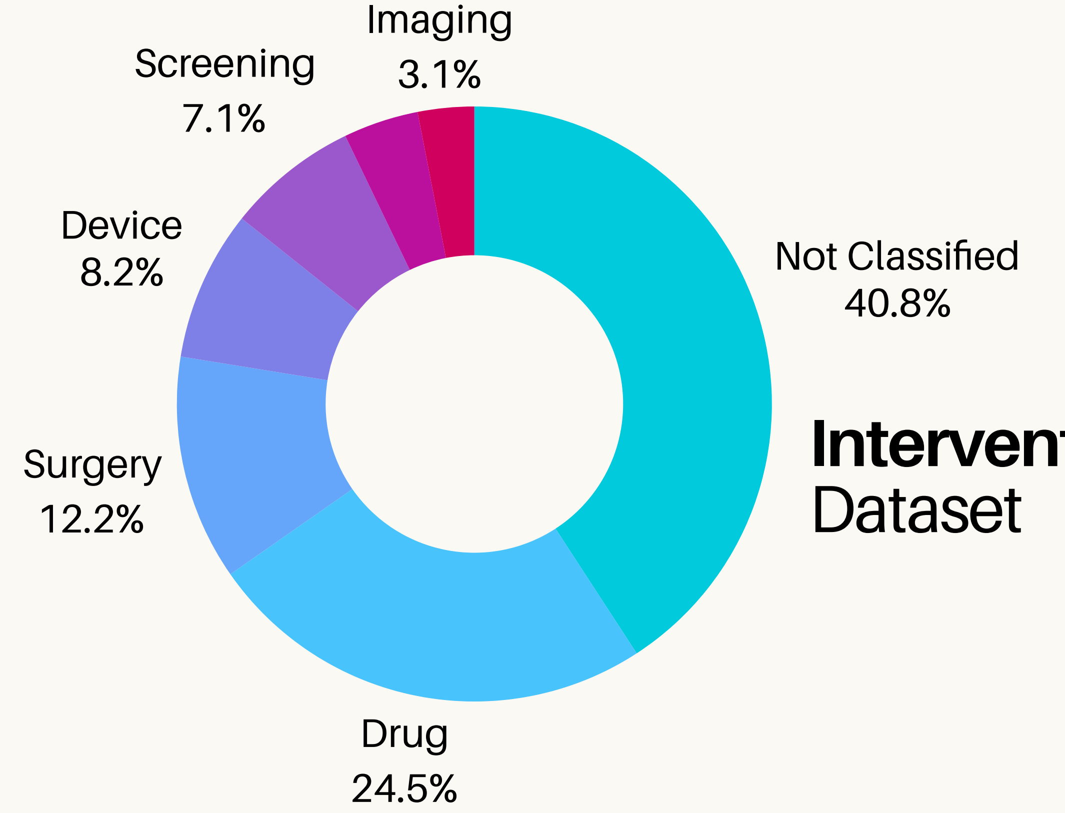
ML Dataset	Mean IR	Max IR	CVIR	Scumble
Clinical Areas	4,53	12,20	1,91	0,63
Interventions	**4,63**	8,82	**2,37**	0,44

Dataset Description

ML Dataset	# samples	Labels	Class Sets	Card	Dens	TCS
Clinical Areas	2192	12	88	0,87	0,07	13,19
interventions	2192	6	23	0,63	0,11	11,12



Clinical Areas Dataset



Interventions Dataset

The **Experiments**

Multi-Label k-Nearest Neighbours

Parameter	Values
K (neighbours)	3, 5, 7, 11
S (smooth)	10, 50, 100, 200

Dataset	K	s
Clinical Areas	11	1
Interventions	11	1

Value-Weighted Multi-Label k-Nearest Neighbours

Parameters	Values
K (neighbours)	1, 3, 5, 7, 11
a	0.3, 0.5, 0.7
b	0.3, 0.5, 0.7

Dataset	K	a	b
Clinical Areas	1	0.5	0.5
Interventions	1	0.3	0.3

The Experiments

RAkEL - Labelset: [2, 3, 4, 5, 6, 7]

Model	Parameter	Values
Gaussian NB	Smoothing	1e-9, 1e-8, 1e-7, 1e-6
Random Forest	N (estimators)	10, 50, 100, 200

Dataset	Labelset	s
Clinical Areas	6	1e-9
Interventions	3	1e-7

Dataset	Labelset	N
Clinical Areas	6	50
Interventions	7	200

Binary Relevance

Model	Parameter	Values
kNN	K (neighbours)	1, 3, 5, 7, 11
Multinomial NB	Alpha	0.3, 0.5, 0.7
SVC	Kernel	linear, rbf, sigmoid
	Penalty Parameter	1e-2, 1e-1, 1e0, 1e1, 1e2

Dataset	K
Clinical Areas	5
Interventions	11

Dataset	Alpha
Clinical Areas	0.3
Interventions	0.3

Dataset	Kernel	C
Clinical Areas	Linear	1e-2
Interventions	Linear	1e-2

The **Multi-Label Classifier** Results

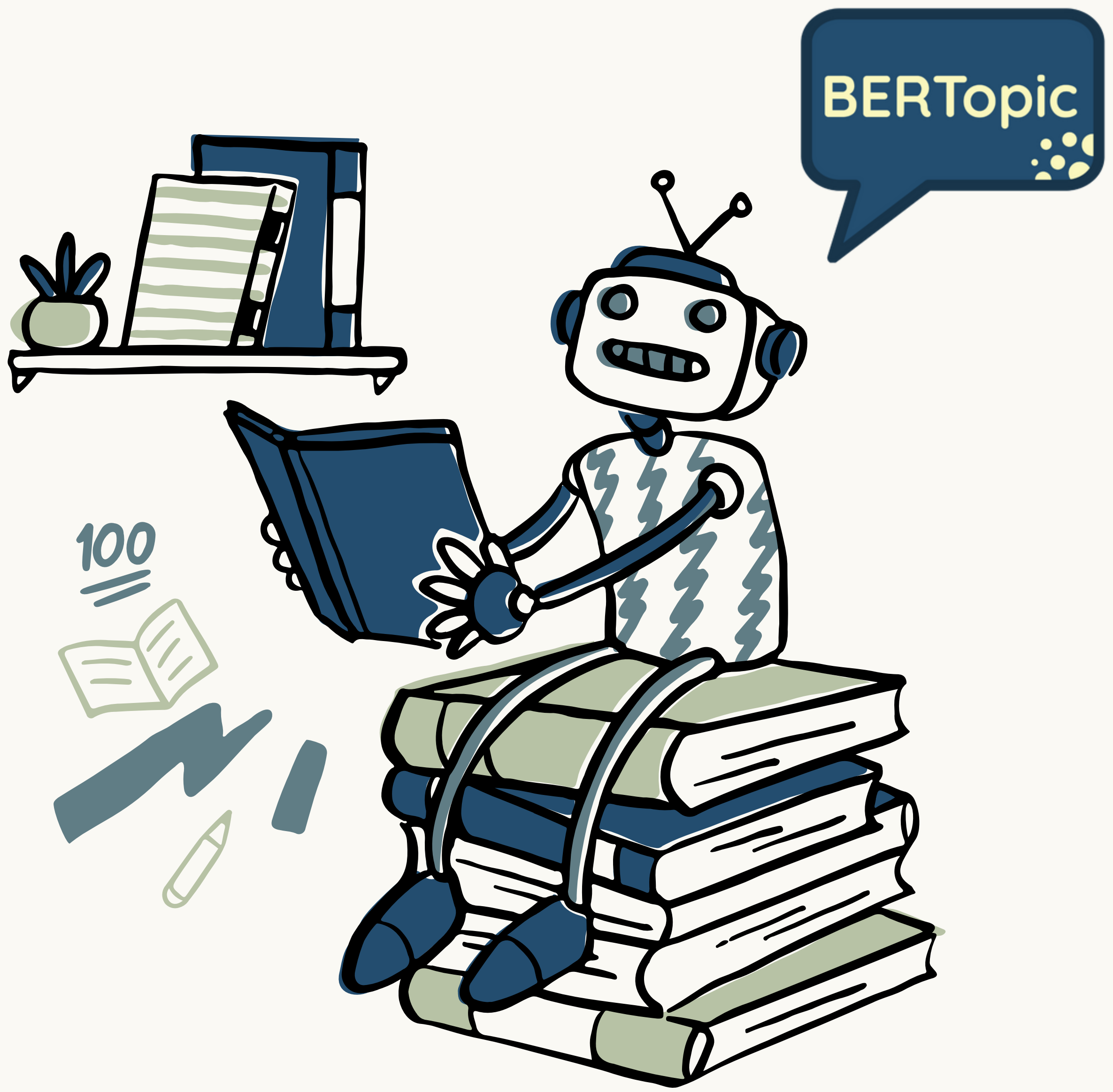
Clinical Areas

Model	F1-Micro	F1-Weighted	Hamming Loss	Ranking Loss	Coverage Error	One Error
ML kNN	0,757	0,748	0,030	0,076	1,714	0,614
VW MLkNN	0,750	0,747	0,034	0,171	2,877	0,470
BR (kNN)	0,778	0,771	0,027	0,200	3,216	0,479
BR (MultinomialNB)	0,704	0,720	0,047	**0,032**	**1,242**	**0,409**
BR (SVC)	**0,823**	**0,815**	**0,022**	0,158	2,735	0,437
RAkEL (Gaussian NB)	0,757	0,759	0,034	0,147	2,698	0,458
RAkEL (Random Forest)	0,426	0,361	0,050	0,510	6,660	0,742

Interventions

Model	F1-Micro	F1-Weighted	Hamming Loss	Ranking Loss	Coverage Error	One Error
ML kNN	0,644	0,633	0,065	0,071	0,984	0,625
VW MLkNN	0,602	0,599	0,083	0,211	1,709	0,643
BR (kNN)	0,648	0,633	0,062	0,228	1,787	0,652
BR (MultinomialNB)	0,630	0,655	0,095	**0,049**	**0,874**	**0,572**
BR (SVC)	**0,710**	**0,694**	**0,053**	0,194	1,648	0,622
RAkEL (Gaussian NB)	0,661	0,675	0,080	0,145	1,396	0,606
RAkEL (Random Forest)	0,571	0,519	0,066	0,298	2,121	0,707

The **BERTopic** Experimental Setup



Binary
Text Classifier
Block

PPS Not Labeled

Clinical Areas:
658 papers

Intervention:
1004 papers

BERTopic Model

- Representation (**KeyBERT**)
- Weight Scheme (**C-TF-IDF**)
- Tokenizer (**CountVectorizer**)
- Clustering (**HDBSCAN**)
- Dim Reduction (**UMAP**)
- Embeddings (**PubMedBERT**)

Model Optimization

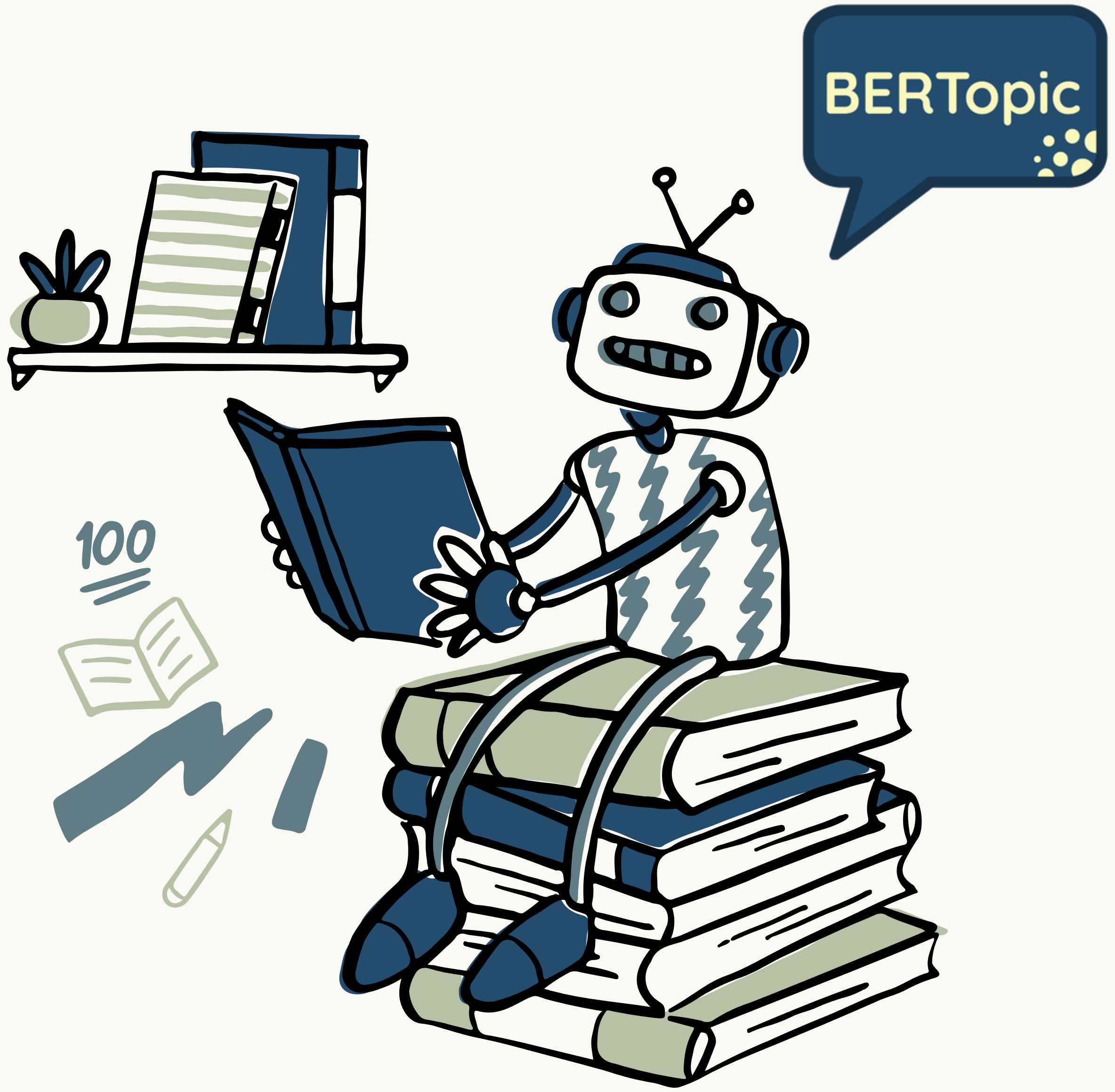
UMAP:
- N_components
- N_neighbours

HDBSCAN:
- Min_clustersize
- Min_samples

Evaluation

- Diversity
- Topics numerosity
- Topics description

The **BERTopic** Results



Data	Topic	Count Papers	Suggested Topic	Label
CLINICAL AREAS	2	36	Psoriasis	**Skin Diseases (NEW)**
	9	21	Eye_cataract_macular	**Visual System Diseases (NEW)**
	15	16	Bladder_oab_urinary	**Urinary Diseases (NEW)**
INTERVENTIONS	13	23	Contracceptive_counseling	**Drugs**
	24	14	Research_testing_diagnonstic	**Screening**

The **Conclusions**

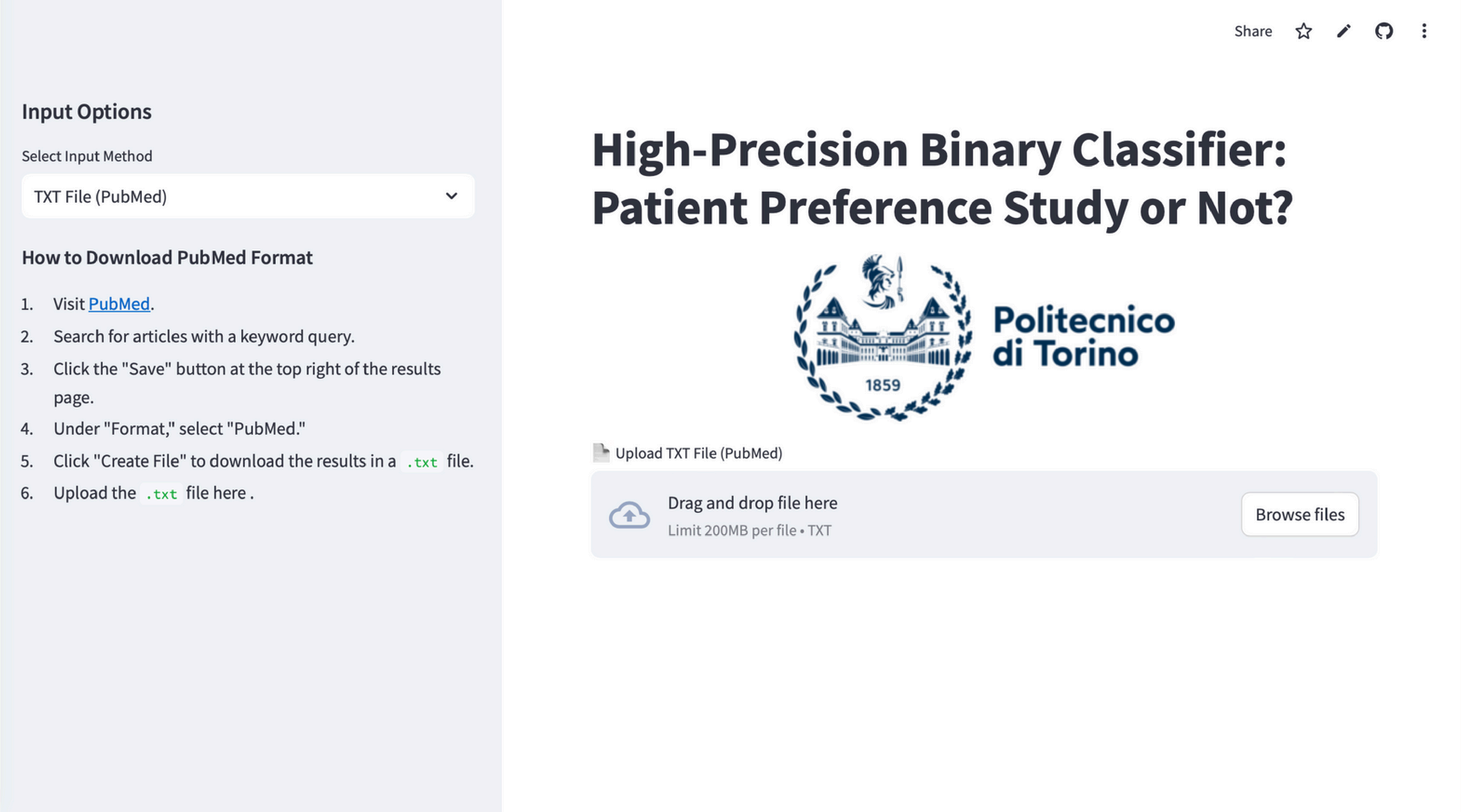
- The **Binary Classifier** performs the first task (PPS identification): built on two ML models (SVC and KNN) on top of two BERT embeddings; a majority voting system drives the class prediction. The model runs with a TPR of 96% and a PPV of 97%.

- The **Multi-Label Classifier** classifies papers of relevance (from the binary classifier) into different categories through a multi-label classification process.

- The investigation explores various multilabel classificators, identifying Binary Relevance with SVC as a classifier, ML kNN, and RAkEL with GaussianNB classifier as effective approaches. Data augmentation serves as the next step to enhance results and address low-frequency label-sets.

- The **BERTopic** study identifies new labels for the clinical areas categorization as well as further items for specific labels in the intervention one.

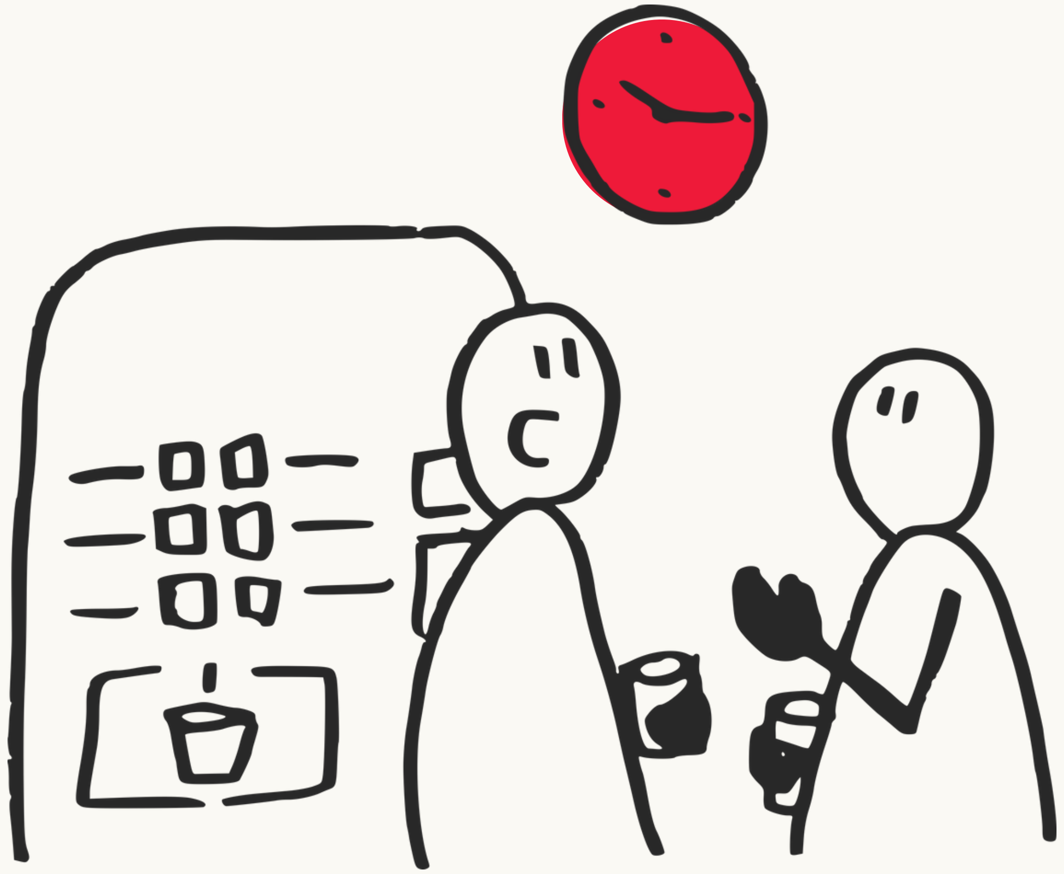
APPLIED DATA SCIENCE PROJECT



https://pps-binary-classifier.streamlit.app

APPLIED DATA SCIENCE PROJECT

Thank You



UNIVERSITÀ DI TORINO

Cesar Augusto Seminario Yrigoyen
Francesco Giuseppe Gillio

Politecnico di Torino