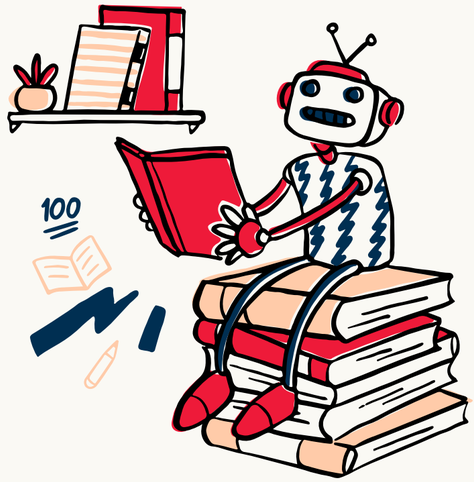
APPLIED DATA SCIENCE PROJECT

Patient Preference Studies
Classification System



Cesar Augusto Seminario Yrigoyen
Francesco Giuseppe Gillio

UNIVERSITÀ DI TORINO

Politecnico di Torino

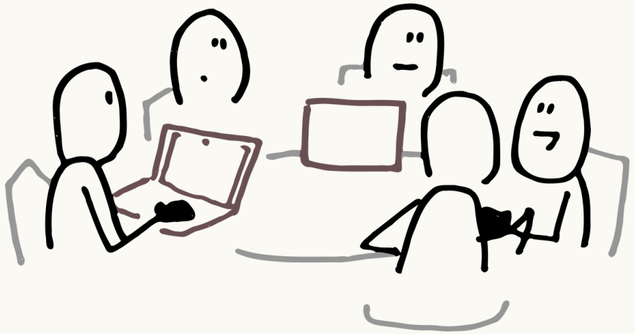
The Picture

European Commission

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

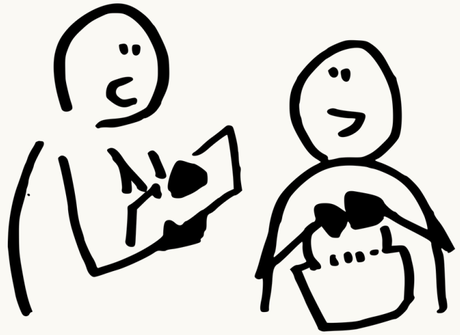
Medical Researchers

actionable insights to



Healthcare Ecosystem

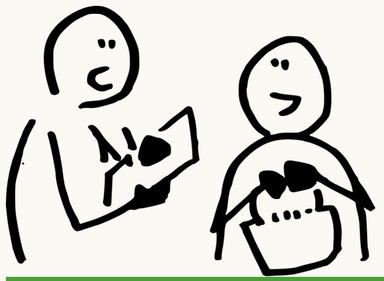
valuable services to



Patient Communities

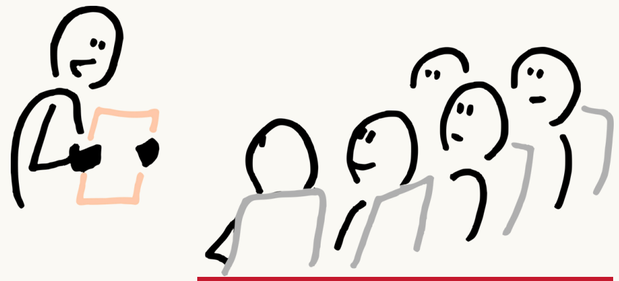
The Values

Sustainable Development Goals
2030



3 GOOD HEALTH AND WELL-BEING

Support the advancement of
patient care systems and processes

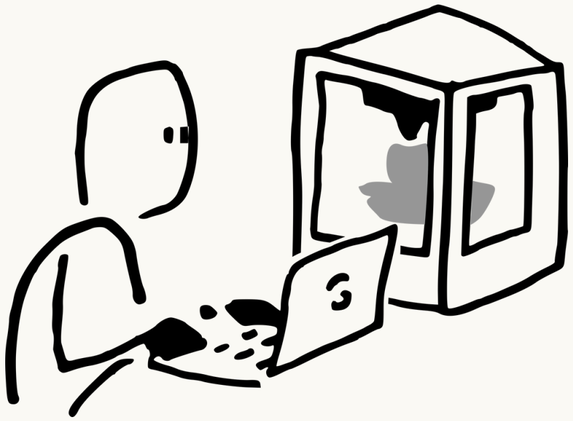


4 QUALITY EDUCATION

Support the advancement of
information retrieval in medical research

The **User** Journey

Data **Collector**



Medical **Researcher**

Data **Source**

Public search engine that provides access to a vast database of references and abstracts on life sciences and biomedical topics, such as research articles and clinical studies

United States National Library of Medicine

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.

The Problem

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

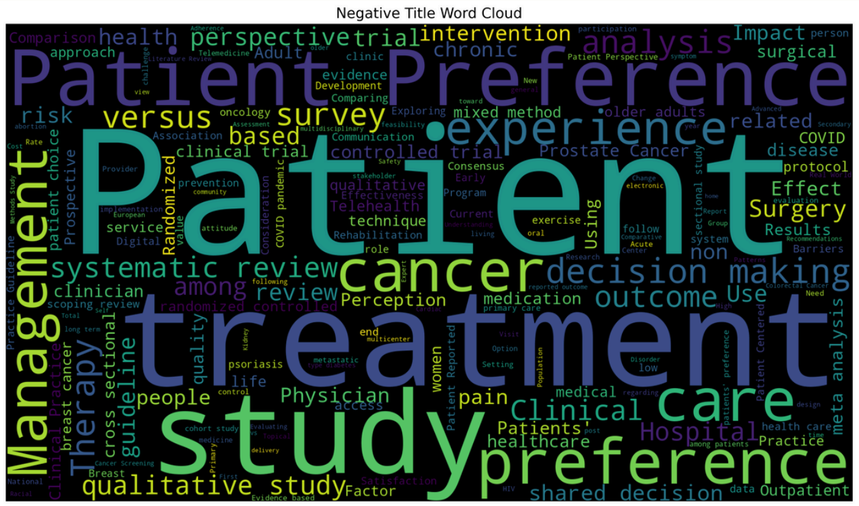
Abstract

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.

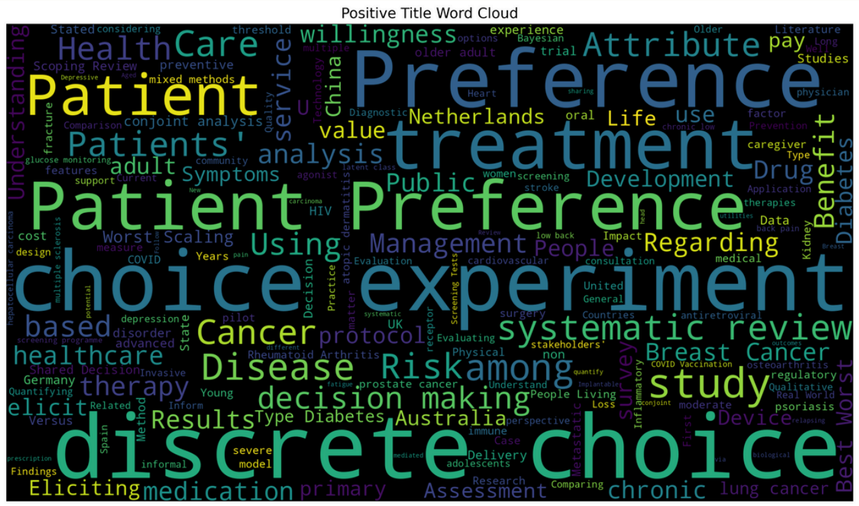
Patient Preference Study or Not?

What about the Clinical Area?

Non Patient Preference Study

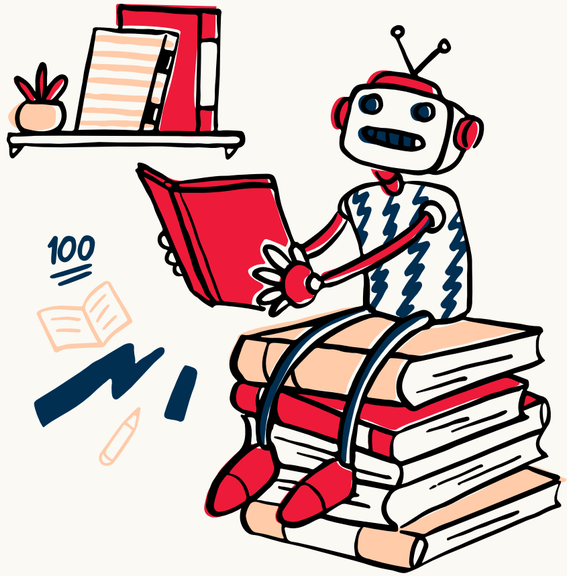


Patient Preference Study



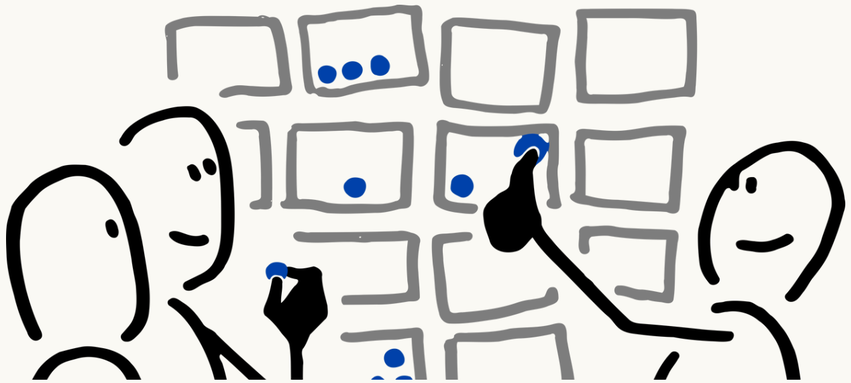
The **Solution**

The non-solution:
read to know



The solution:
teach to read

The Objective



Classifier Model for Medical Research Papers

- Papers classification by relevance to Patient Preference Studies (PPS)
- Papers classification by relevance to Clinical Areas

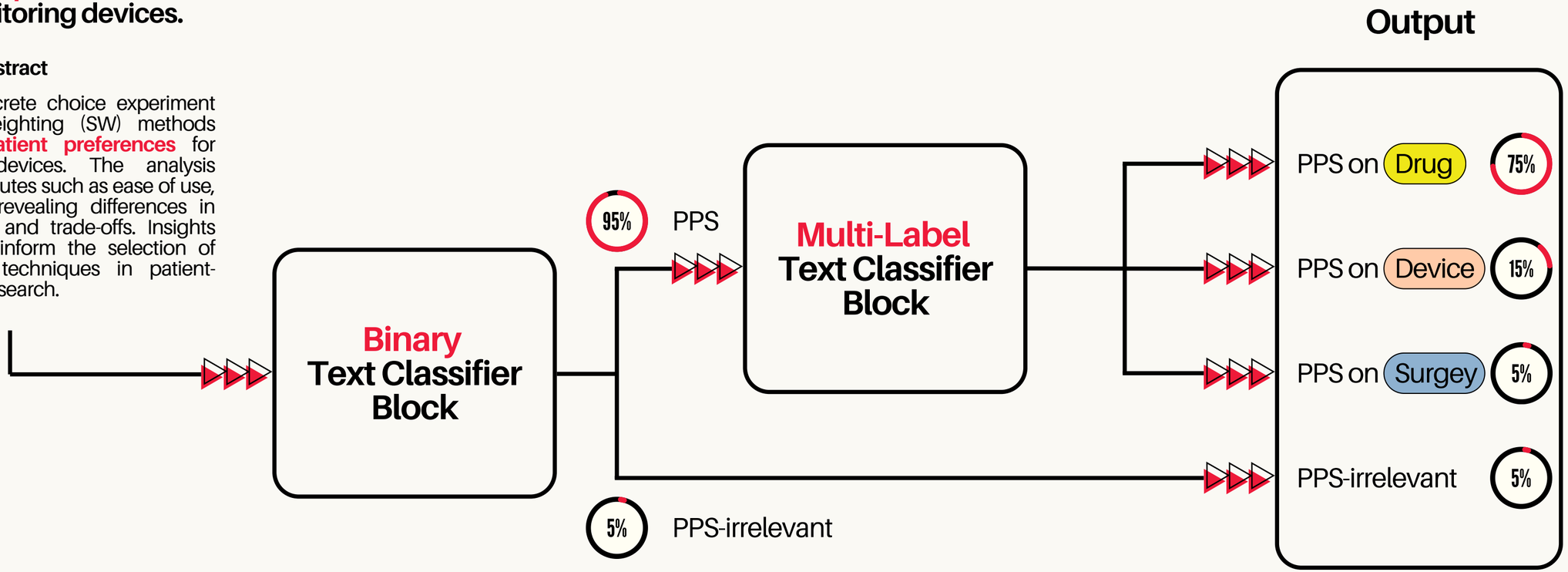
The Task

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.

Two-Stage Text Classification



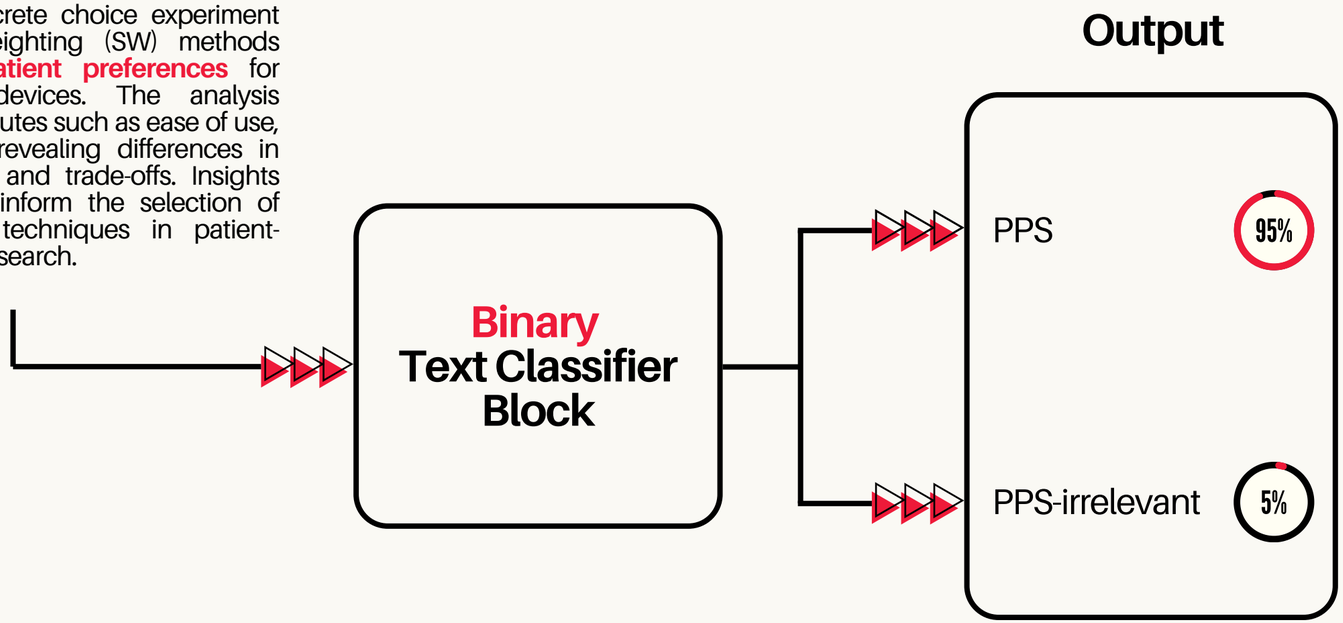
The **Binary** Text Classifier

Classification by relevance to Patient Preference Studies

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.

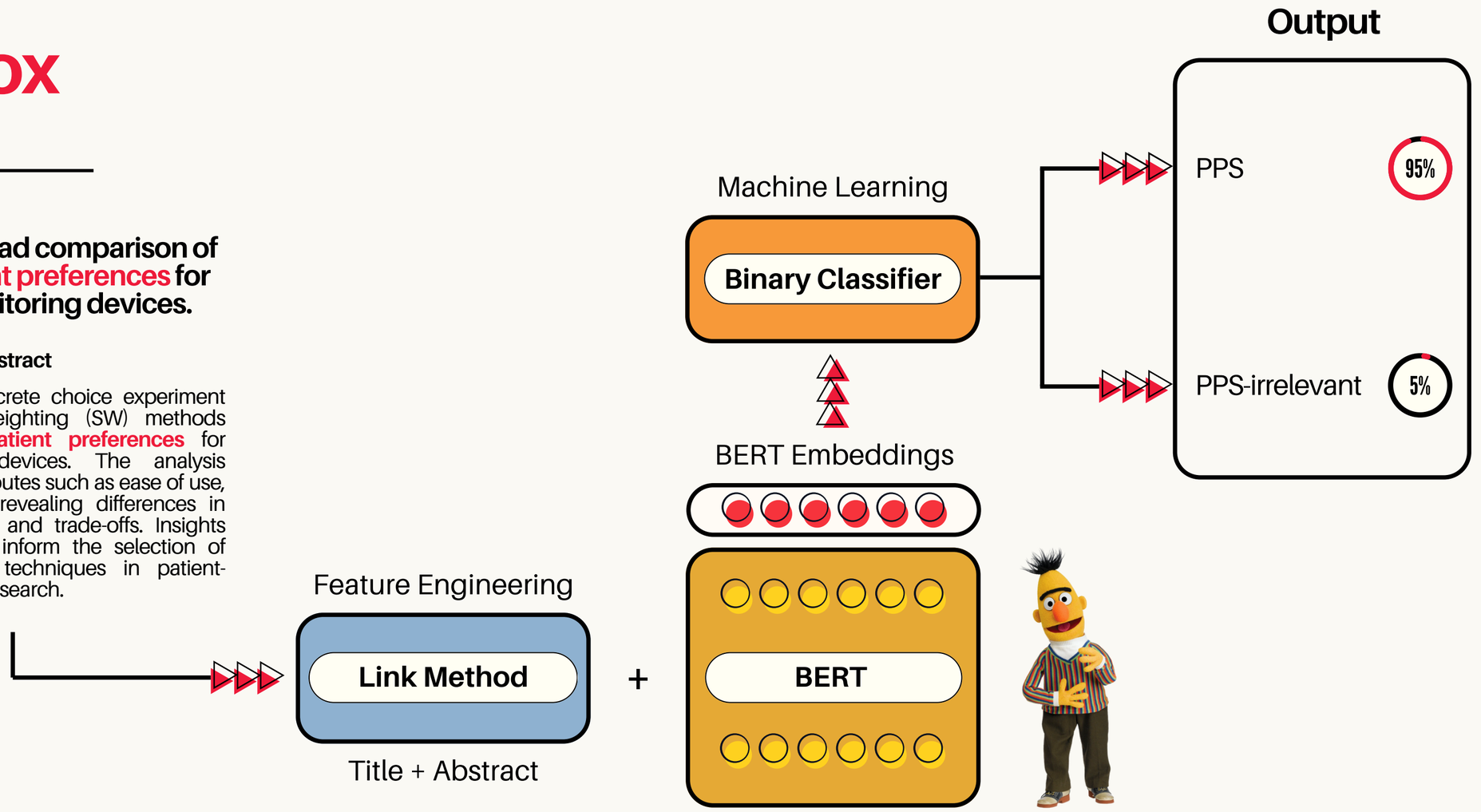


Inside
The Box

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 **Evaluation** Metrics

What to evaluate?

The worst outcome:
researchers lose papers on PPS

The target outcome:
researchers retrieve papers on PPS alone

 remove noise

True Positive Rate (TPR)

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

The probability that an actual positive tests positive

Positive Predictive Value (PPV)

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

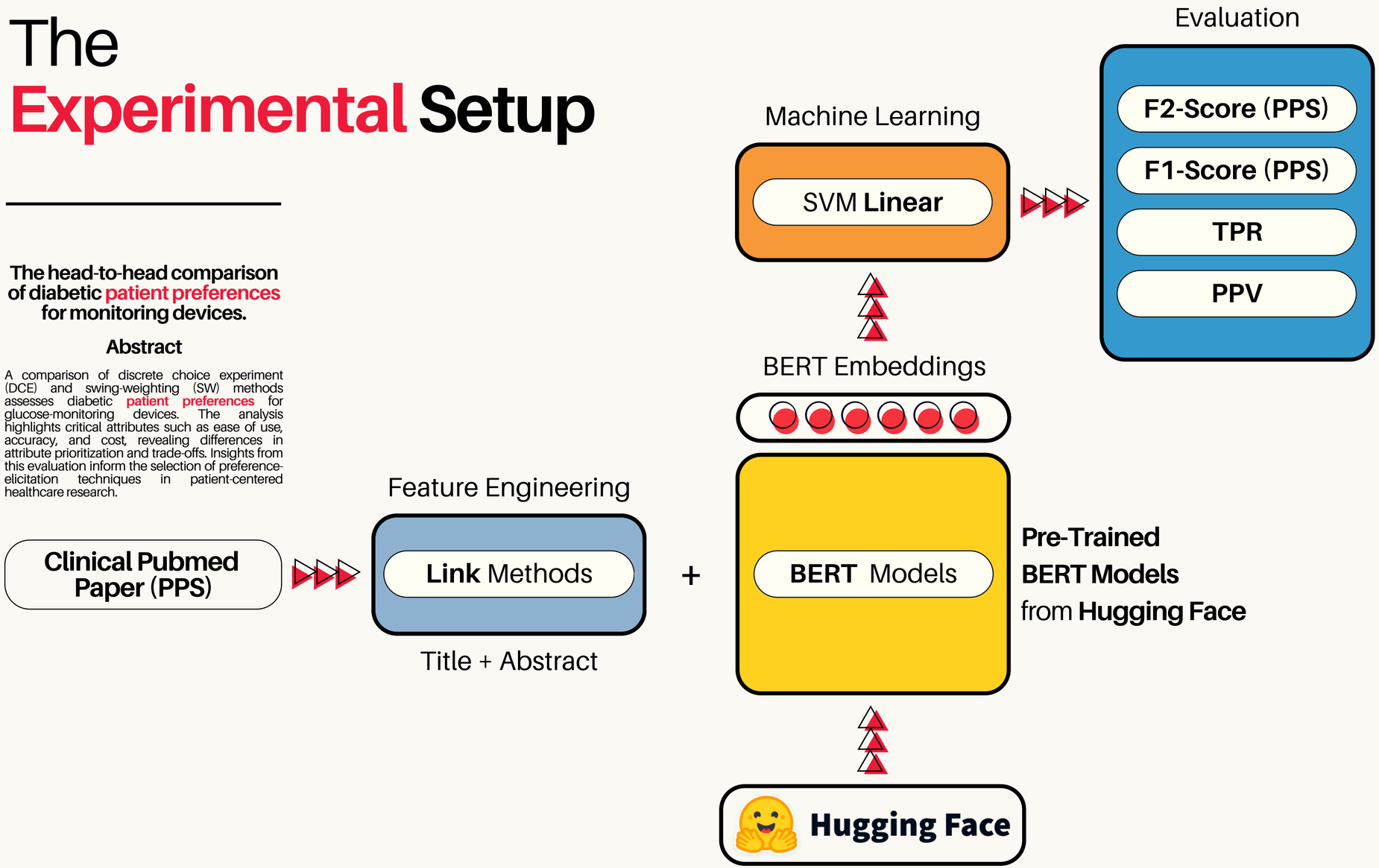
The probability that a positive test matches an actual positive

The Experimental Setup

The head-to-head comparison of diabetic patient preferences for 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.

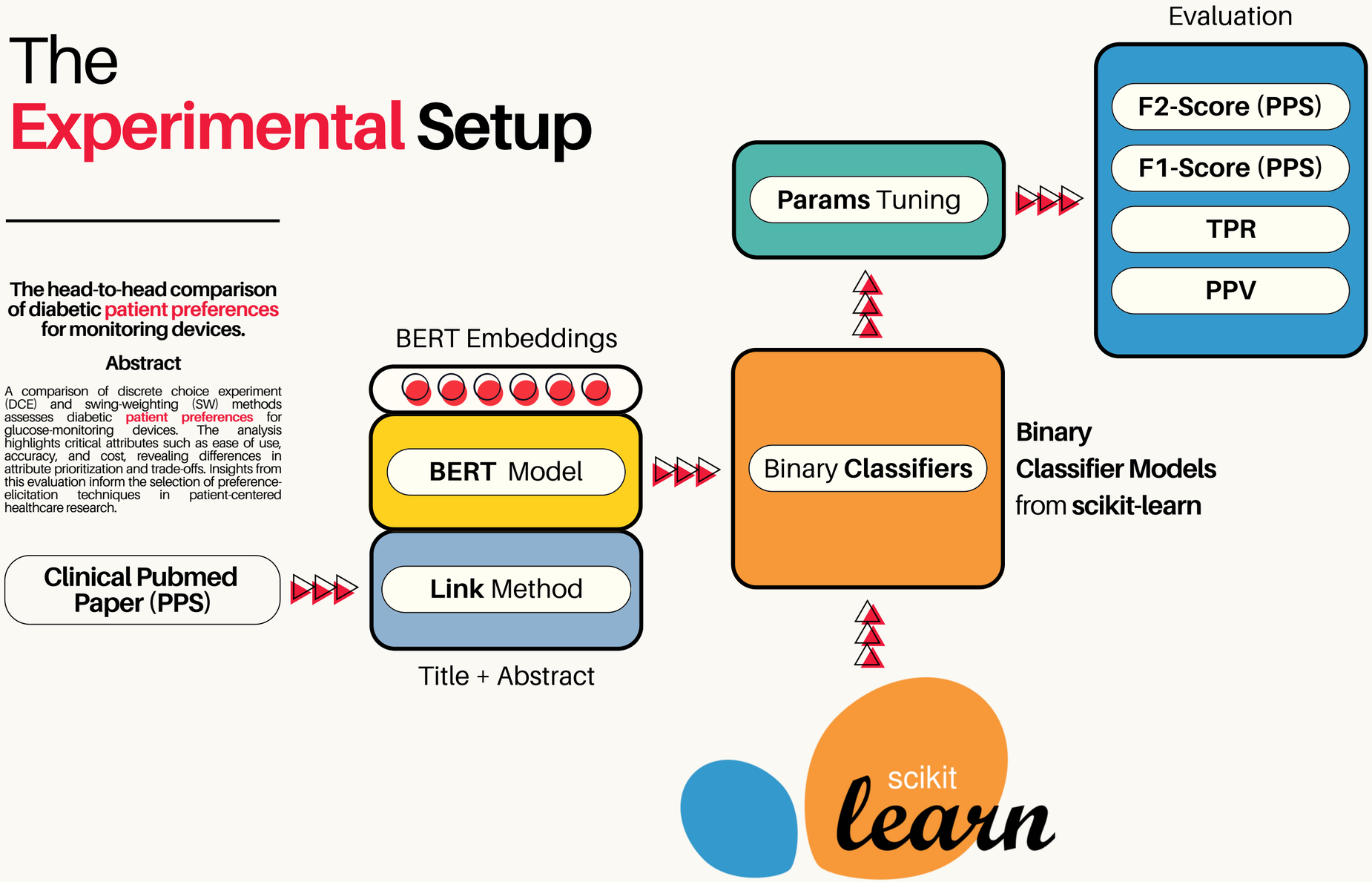


The Experimental Setup

The head-to-head comparison of diabetic patient preferences for 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

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

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**

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**

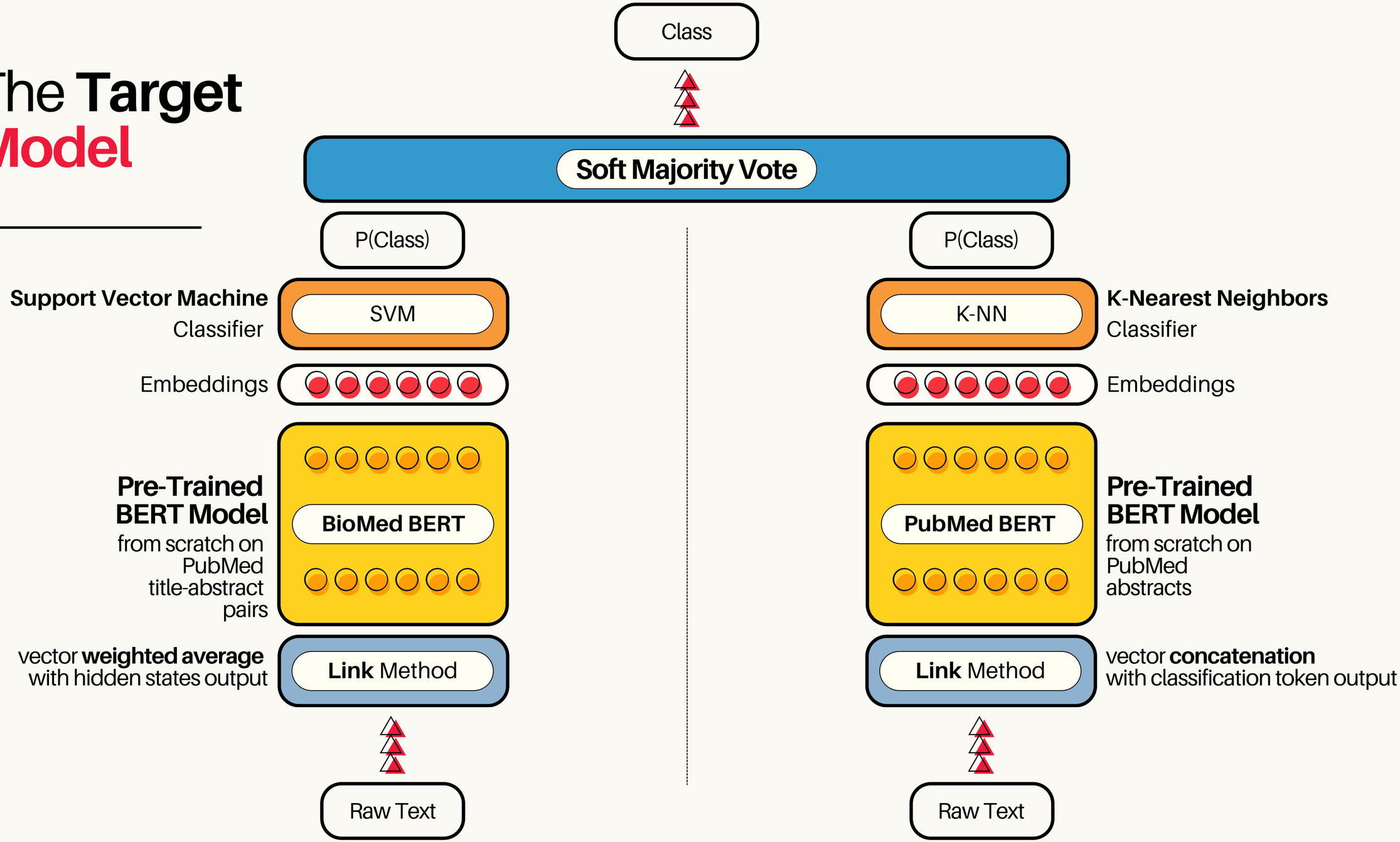
High Recall Model

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

High Precision Model

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

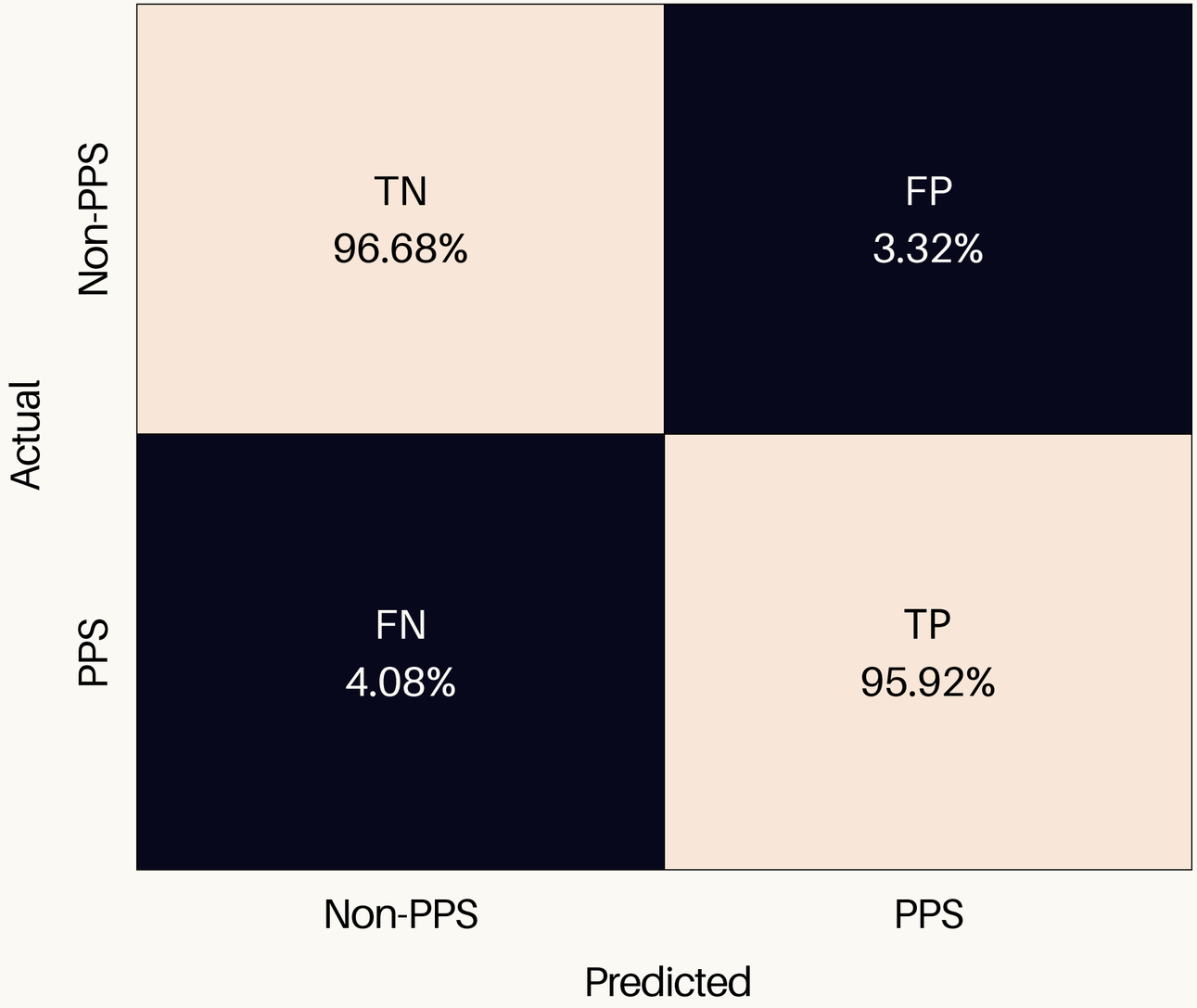
The **Target** Model



The **Result**

PubMed-BERT and **K-NN** x **0.4375**
+
BioMed-BERT and **SVM** x **0.5625**

with **threshold** = **0.3875**



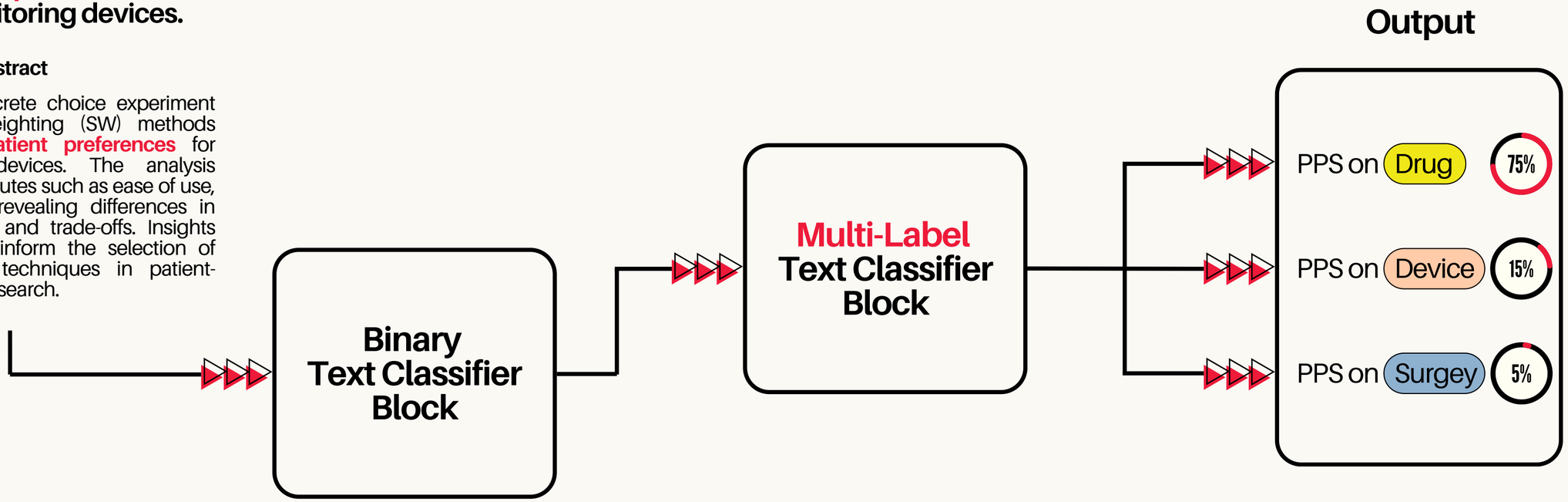
The **Multi-Label** Text Classifier

Classification by relevance to Clinical Areas

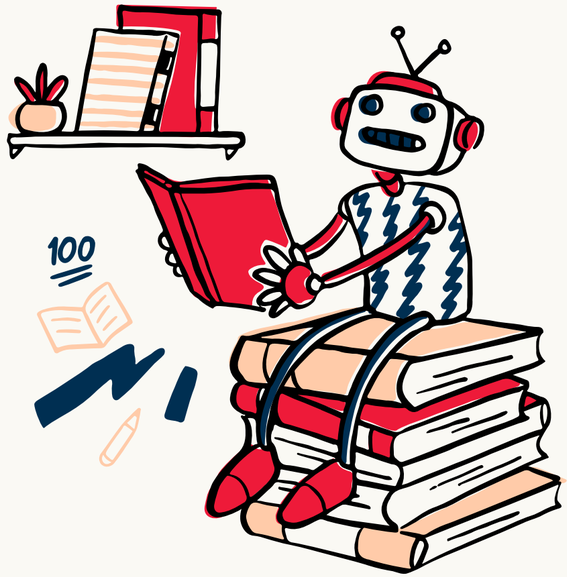
The head-to-head comparison of diabetic patient preferences for glucose-monitoring devices.

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

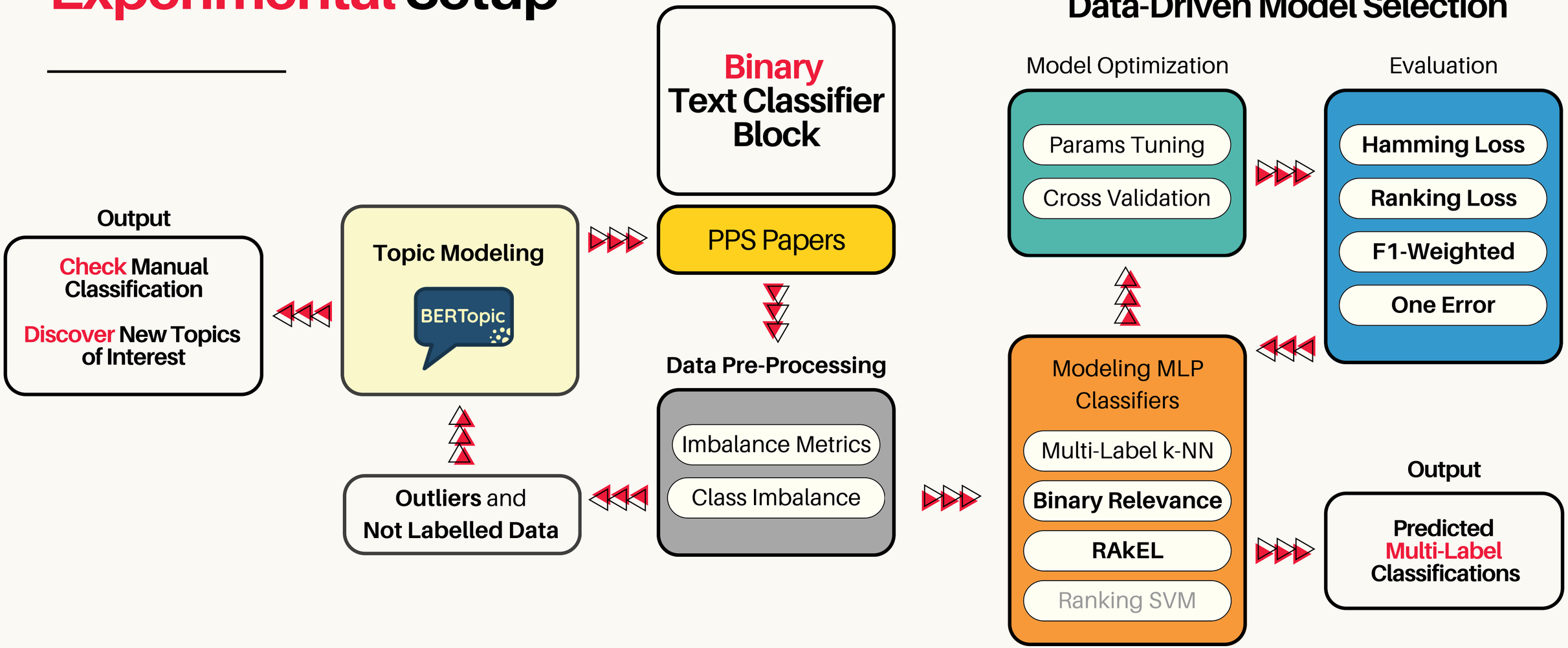
- Problem transformation (Binary relevance, RAkEL)
- Algorithm adaptation (multilabel kNN, VW ML kNN, ranking SVM)
- Ensemble models
- Cost sensitive (RAkEL)

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

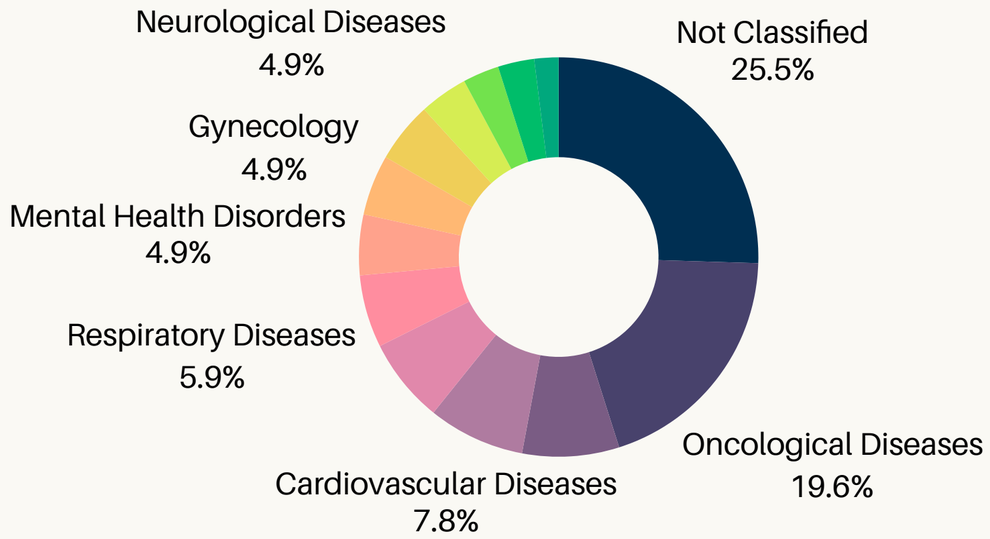
Dataset Imbalance

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

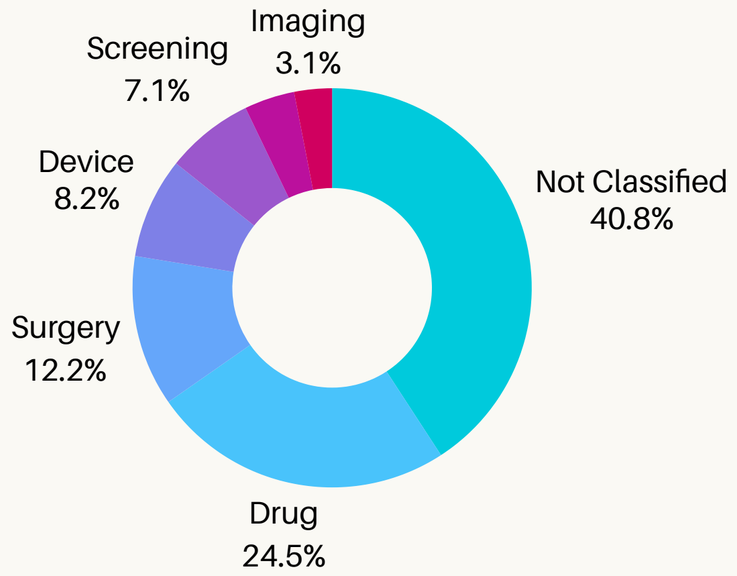
Dataset Description

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

Parameters	Values
K (neighbours)	3, 5, 7, 11
S (smooth)	0.5, 1

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 Results**

Final BERTopic Configuration

Representation Tuning	KeyBERT, LLama
Weighted Scheme	C-TF-IDF
Tokenizer	CountVect
Clustering	HDBSCAN
Dimensionality Reduction	UMAP
Embeddings	SBERT (Pubmedbert)

Applied on **“Not Labelled Data”**

	Topic	Count	Name	Representation	KeyBERT	MMR	POS	Representative_Docs
0	-1	290	-1_service_medical_used_data	[service, medical, used, data, time, research,...	[respondent, utility, measure, attribute, cost...	[service, medical, used, data, time, research,...	[service, medical, data, time, research, attri...	[latent class model heterogeneity latent class...
1	0	148	0_family_endoflife_home_caregiver	[family, endoflife, home, caregiver, advance, ...	[nursing home, end life, advance planning, eld...	[family, endoflife, home, caregiver, advance, ...	[family, endoflife, home, caregiver, advance, ...	[unpacking impact adult home death family care...
2	1	123	1_cancer_breast_information_breast cancer	[cancer, breast, information, breast cancer, o...	[breast cancer, cancer survivor, lung cancer, ...	[cancer, breast, information, breast cancer, o...	[cancer, breast, information, oncology, role, ...	[understanding value regarding early stage lun...
3	2	85	2_attribute_method_dce_healthcare	[attribute, method, dce, healthcare, data, exp...	[experiment dces, dces, technology assessment,...	[attribute, method, dce, healthcare, data, exp...	[attribute, method, healthcare, data, experime...	[novel design process selection attribute incl...
4	3	63	3_colleague_lesson_routine practice_say	[colleague, lesson, routine practice, say, nh,...	[, , , , , , , ,]	[colleague, lesson, routine practice, say, nh,...	[lesson, routine practice, right, shift, progr...	[, , implementing nh lesson magic programme ...
5	4	51	4_sdm_clinician_practice_physician	[sdm, clinician, practice, physician, option, ...	[sdm, practice, clinician, physician, provider...	[sdm, clinician, practice, physician, option, ...	[sdm, clinician, practice, physician, option, ...	[assessing option gridxae practicability feasi...
6	5	47	5_woman_pregnancy_contraceptive_attribute	[woman, pregnancy, contraceptive, attribute, m...	[pregnancy, woman, mother, fertility, pregnant...	[woman, pregnancy, contraceptive, attribute, m...	[woman, pregnancy, contraceptive, attribute, m...	[woman attribute firsttrimester miscarriage ma...
7	6	45	6_tto_state_utility_value	[tto, state, utility, value, time, tradeoff, v...	[state valuation, utility value, valuation, st...	[tto, state, utility, value, time, tradeoff, v...	[tto, state, utility, value, time, tradeoff, v...	[correcting value influence importance correct...
8	7	43	7_mental_depression_service_sdm	[mental, depression, service, sdm, consumer, s...	[mental, schizophrenia, depression, sdm, psych...	[mental, depression, service, sdm, consumer, s...	[mental, depression, service, sdm, consumer, u...	[family involvement consumer serious mental il...
9	8	34	8_prostate_prostate cancer_cancer_men	[prostate, prostate cancer, cancer, men, decis...	[prostate cancer, prostate, localized prostate...	[prostate, prostate cancer, cancer, men, decis...	[prostate, cancer, men, decisional, da, decisi...	[voice methodology novel mixedmethods approach...
10	9	30	9_pain_exercise_low_effect	[pain, exercise, low, effect, participation, d...	[pain, level, effectiveness, musculoskeletal, ...	[pain, exercise, low, effect, participation, d...	[pain, exercise, low, effect, participation, d...	[people considering exercise prevent low back ...
11	10	26	10_diabetes_type diabetes_type_utility	[diabetes, type diabetes, type, utility, le, s...	[diabetes, type diabetes, diagnosis diabetes, ...	[diabetes, type diabetes, type, utility, le, s...	[diabetes, type, utility, state, program, beha...	[young adult type diabetes clinic approach att...
12	11	21	11_dental_wtp_oral_dentist	[dental, wtp, oral, dentist, role, preferred	[dental, dentist, teeth, willingnesstopay, pay	[dental, wtp, oral, dentist, role, preferred	[dental, wtp, oral, dentist, role, preferred	[preferred perceived control dental increasing...

The **Conclusions**

- The **binary classifier** performs the first task (PPS identification): built on 2 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 0.96 and a PPV of 0.97.

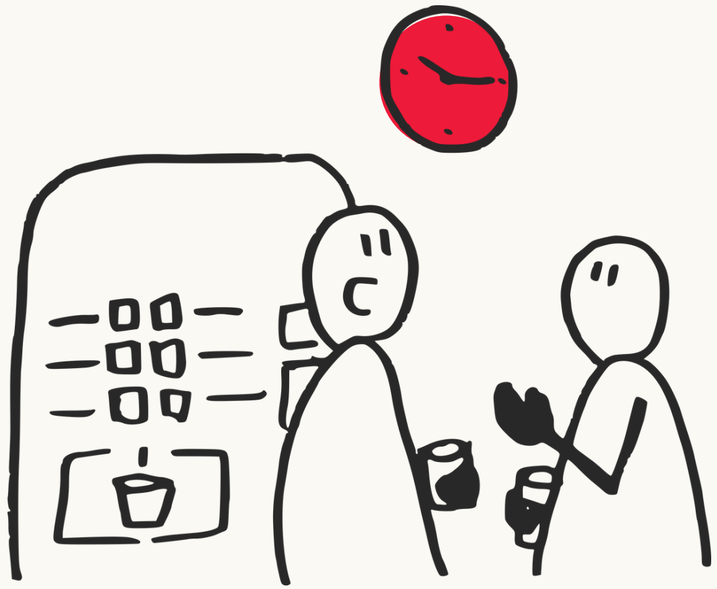
- The **multi-label classifier** classifies papers of relevance (from the binary classifier) into different categories through a multilabel classification process.

- 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 a new label, diabetes, within the clinical areas category and proposes additional topics for further categorization.

APPLIED DATA SCIENCE PROJECT

Thank You



Cesar Augusto Seminario Yrigoyen
Francesco Giuseppe Gillio

UNIVERSITÀ DI TORINO

Politecnico di Torino