

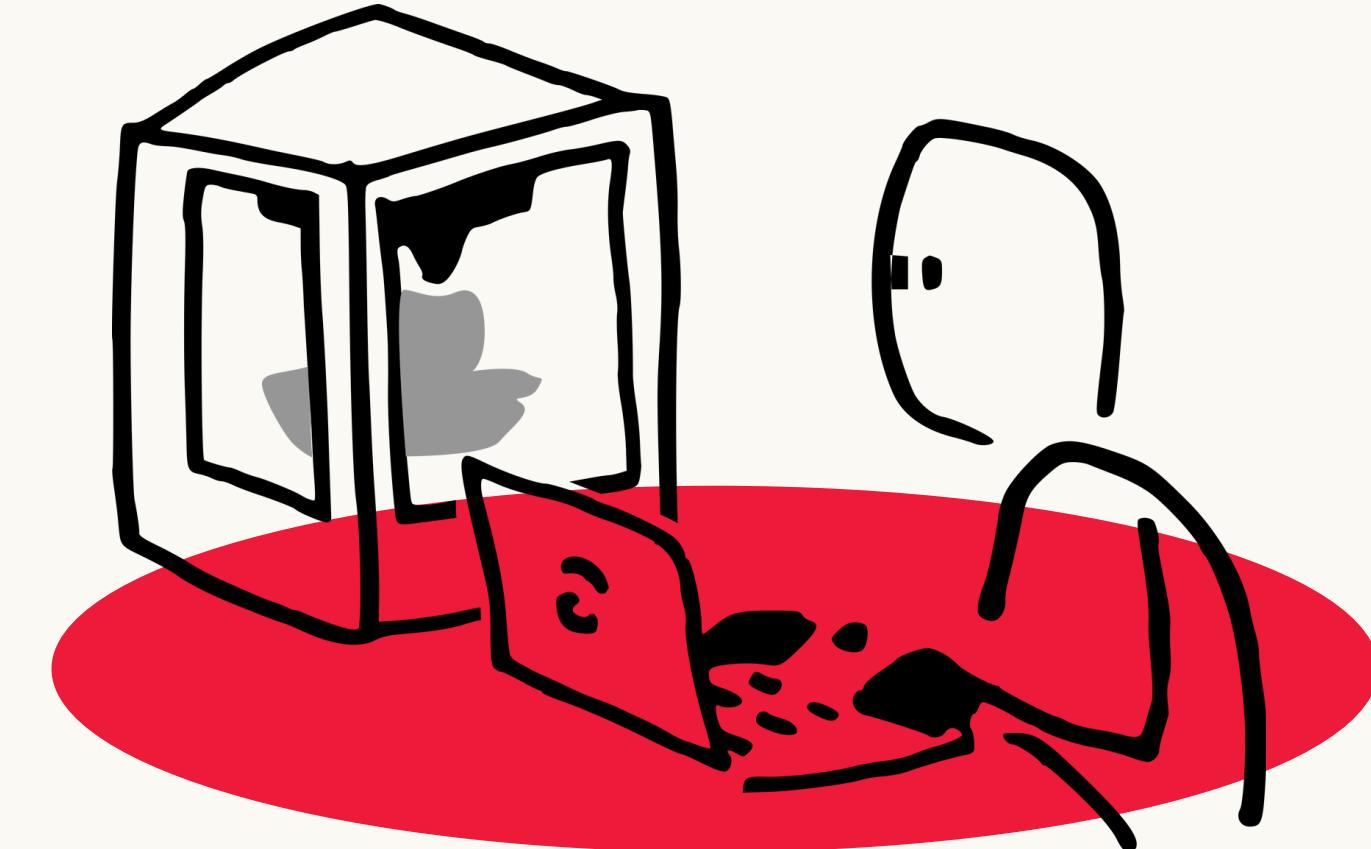
APPLIED DATA SCIENCE PROJECT

Patient Preference Studies Classification System

Francesco Giuseppe Gillio & Cesar Augusto Seminario Yrigoyen



UNIVERSITÀ
DI TORINO



Politecnico
di Torino

APPLIED DATA SCIENCE PROJECT



Checkpoint



UNIVERSITÀ
DI TORINO



Politecnico
di Torino

Table of Contents

Part 1

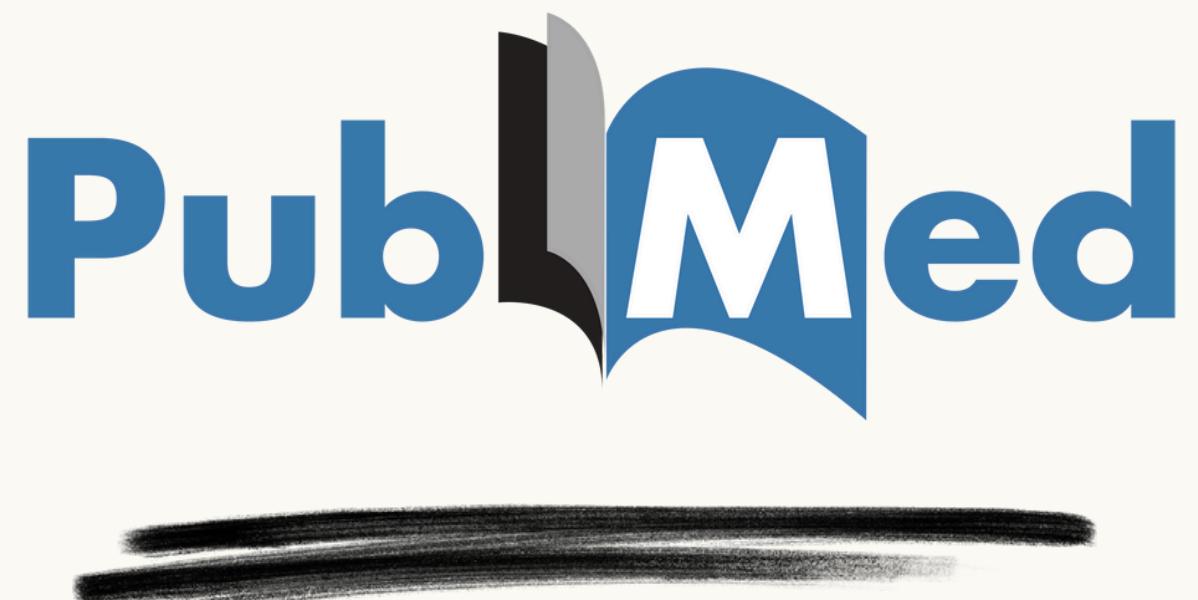
- ▶▶▶ Project **Background**
- ▶▶▶ Project **Value Proposition**
- ▶▶▶ Project **General Objectives**
- ▶▶▶ Project **Design**
- ▶▶▶ Project **Work Breakdown Structure**

The Data

Patient Preference Studies

Patient Preference Studies explore
therapy attributes important to **patient communities**,
their significance, and acceptable trade-offs

Available on



UNIVERSITÀ
DI TORINO



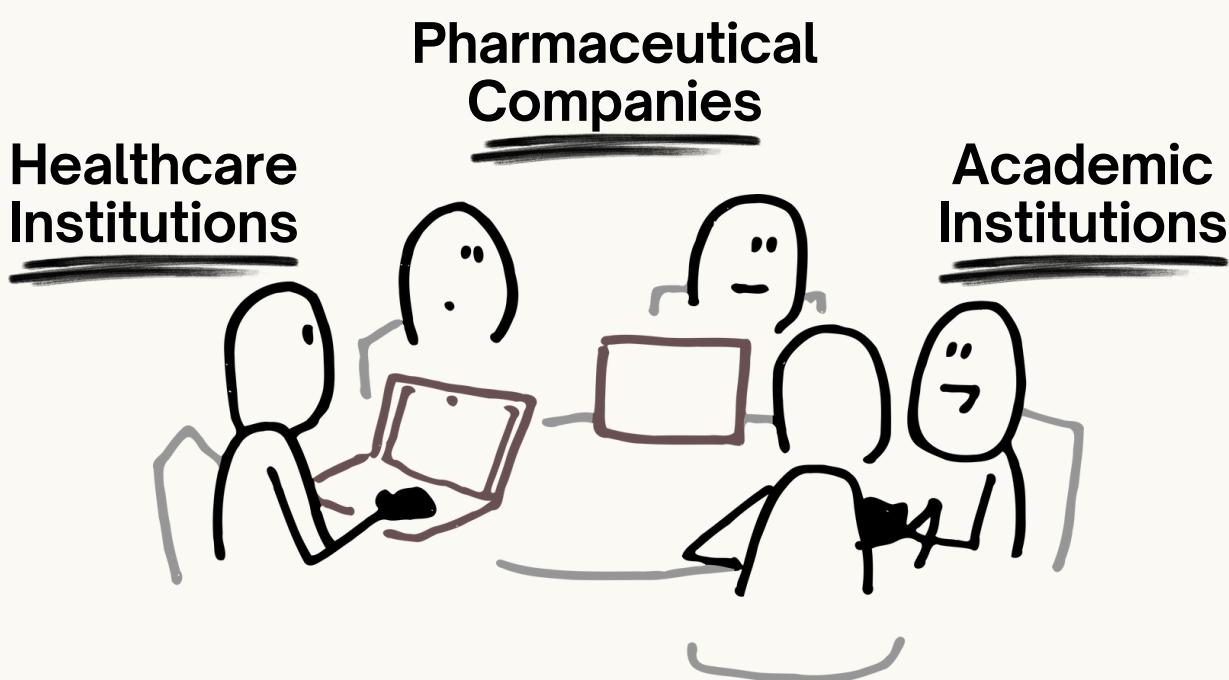
Politecnico
di Torino

The Stakeholders



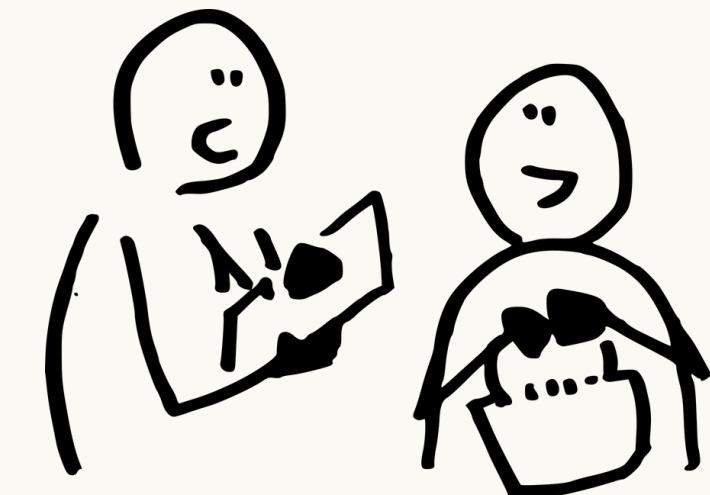
Medical
Researchers

CORE



Healthcare
Ecosystem

DIRECT



Patient
Communities

INDIRECT

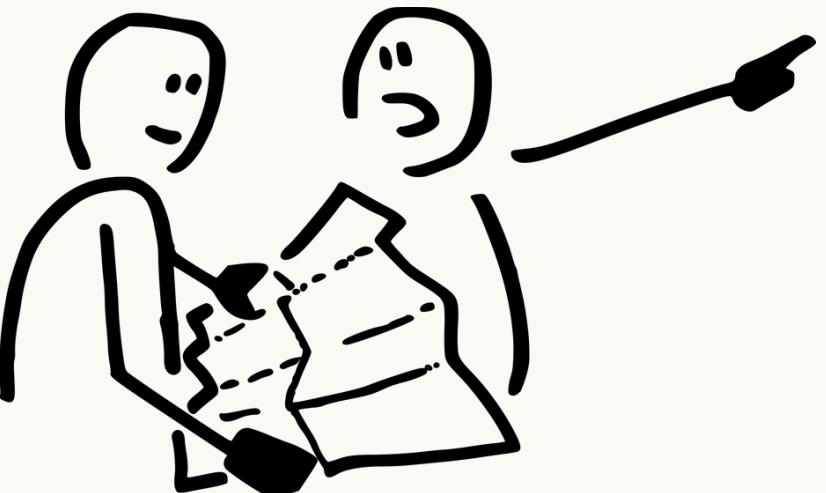
The Challenges



Large Volume

of Scientific Literature

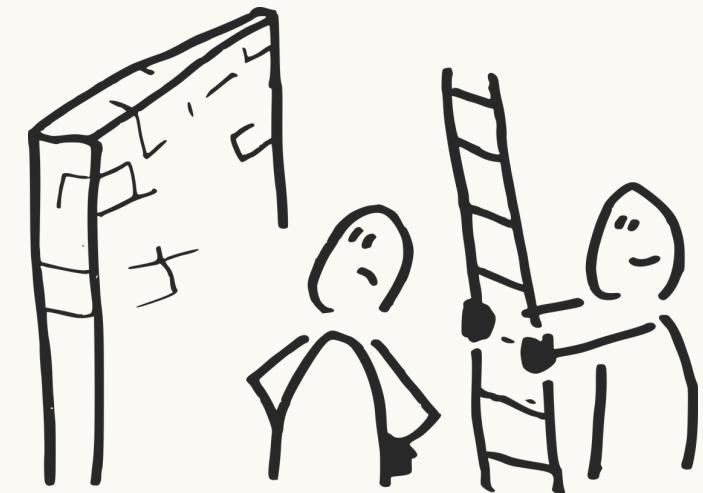
search strings in citation databases (**PubMed**) return a large amount of content, often irrelevant



Broad Scope

of Scientific Areas

PPS cover a wide range of clinical areas and accurate searches require manual supervision

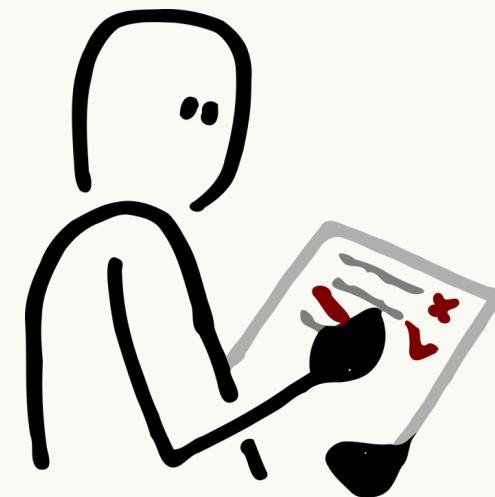


Adaptation to Scale

of Scientific Databases

manual supervision struggles to cope with the publication scale of scientific literature

The Project Values



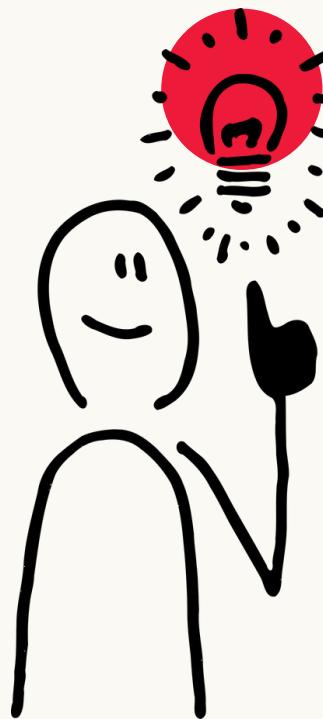
Improve Relevance

by **Classification System**
to bypass irrelevant content and
return high-value literature



Improve Retrieval

by **Categorization System**
to categorize search results and
improve area-specific retrieval



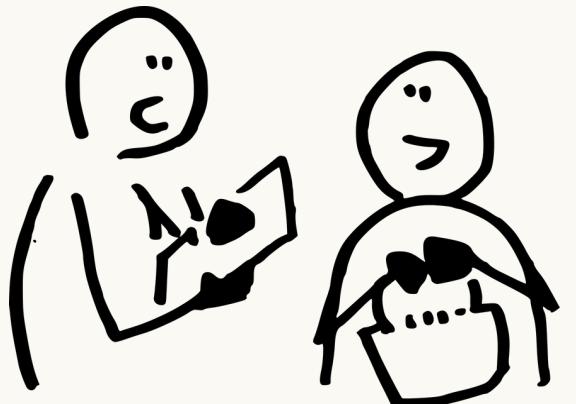
Improve Efficiency

by **System Automation**
to reduce manual effort and improve
access to up-to-date research

The Project Values



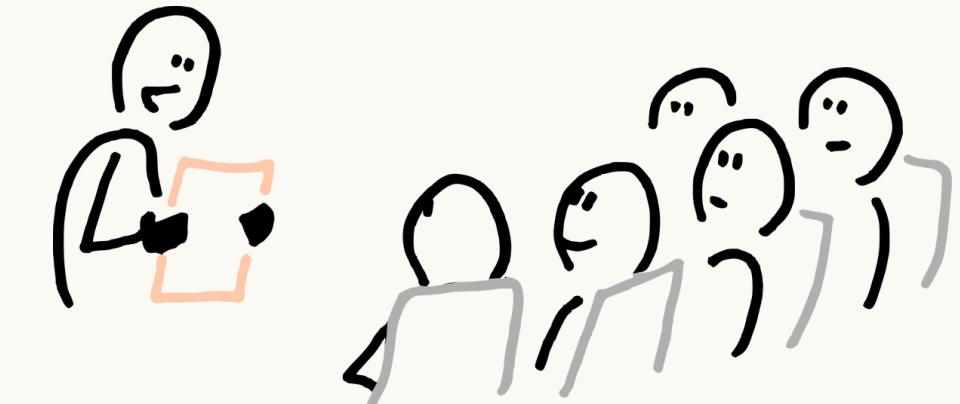
Sustainable Development Goals Alignment



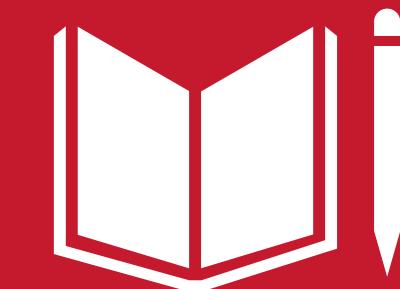
3 GOOD HEALTH AND WELL-BEING



Support the advancement of **patient care** through accurate and high-quality patient-relevant data

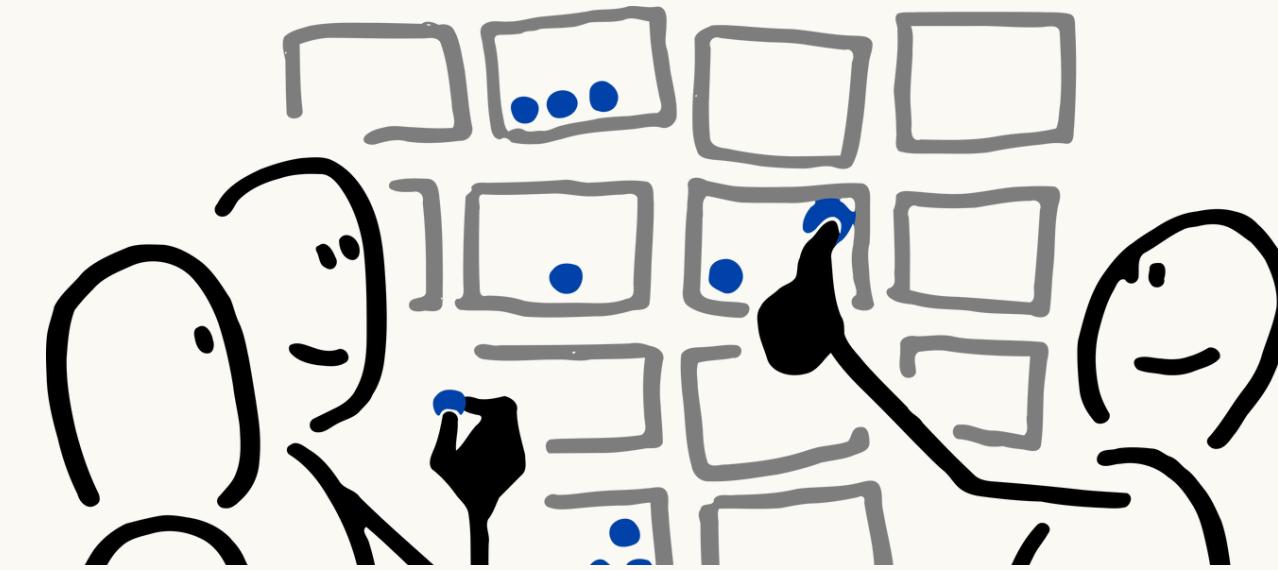


4 QUALITY EDUCATION



Improve information retrieval in **medical and health research** through accurate, high-quality and relevant literature

The **Project Objective**



Classifier Model for Scientific Papers



Detect Patient Preference Studies from PubMed outputs



Categorize PPS-relevant papers into **Clinical Areas**

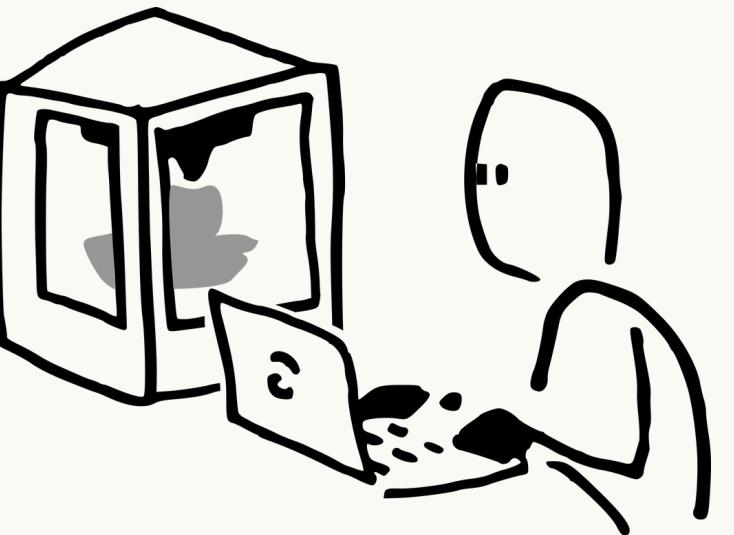
by **Titles and Abstracts**

The **Project Objective**

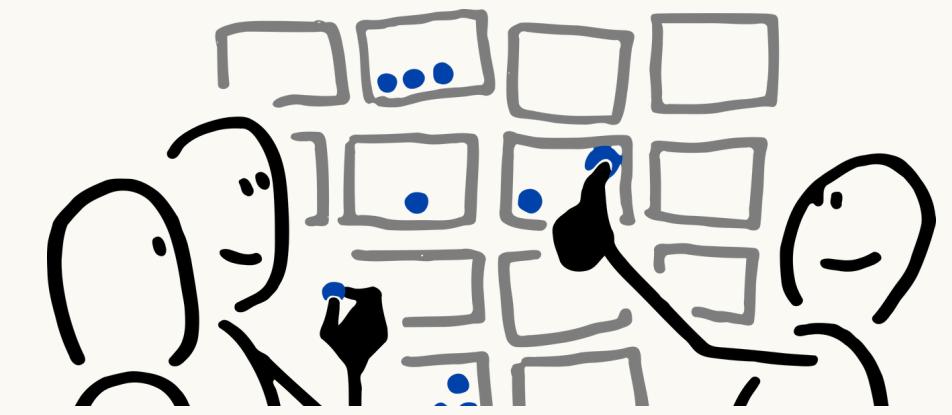


**Supervised
Binary
Classifier Model**

to classify
search string outputs
by relevance to **PPS**



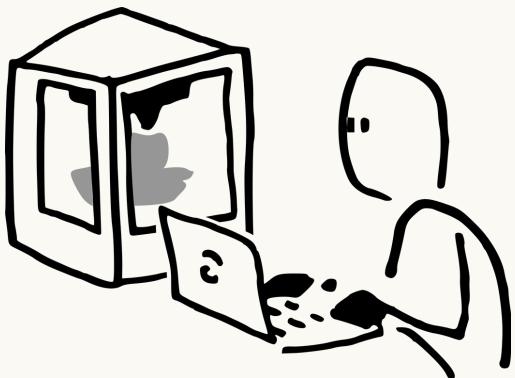
Titles and Abstracts



**Self-Supervised
Multi-Label
Classifier Model**

to categorize
PPS-relevant content
into **Clinical Areas**

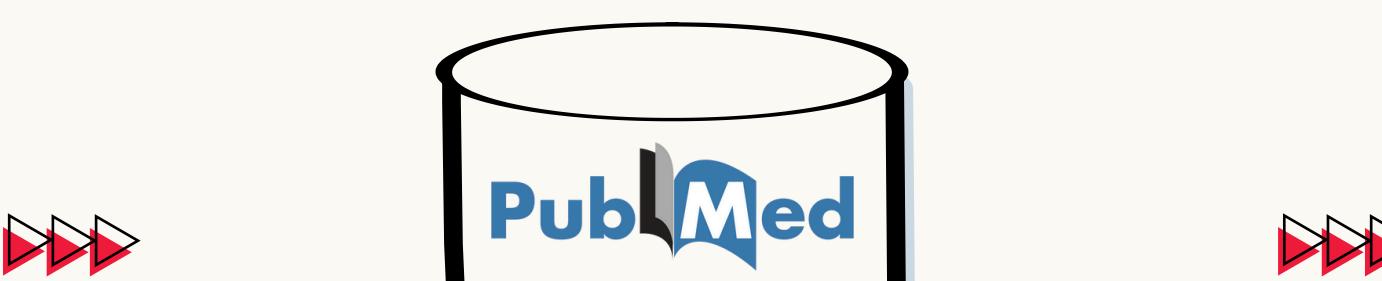
The Project Design



Medical Researcher



High-quality PPS-relevant literature

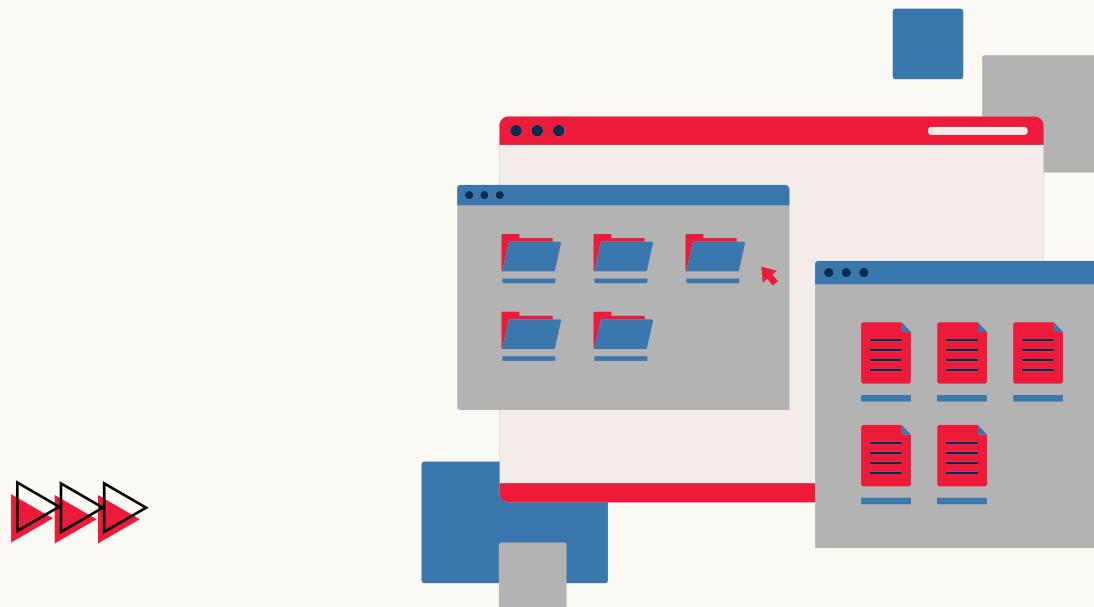


Large Search String on
Patient Preference Studies

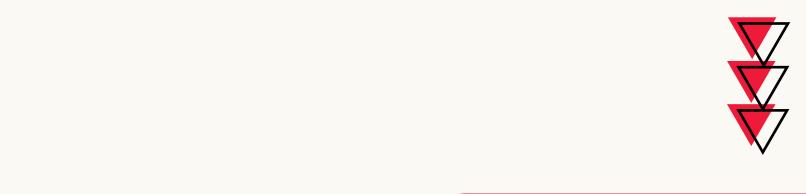


**Self-Supervised
Multi-Label
Classifier Model**

to categorize PPS-relevant
content into Clinical Areas



PPS-relevant and irrelevant
content across clinical
areas



**Supervised
Binary
Classifier Model**

to classify search string
outputs by relevance to PPS

Project Work Breakdown Structure

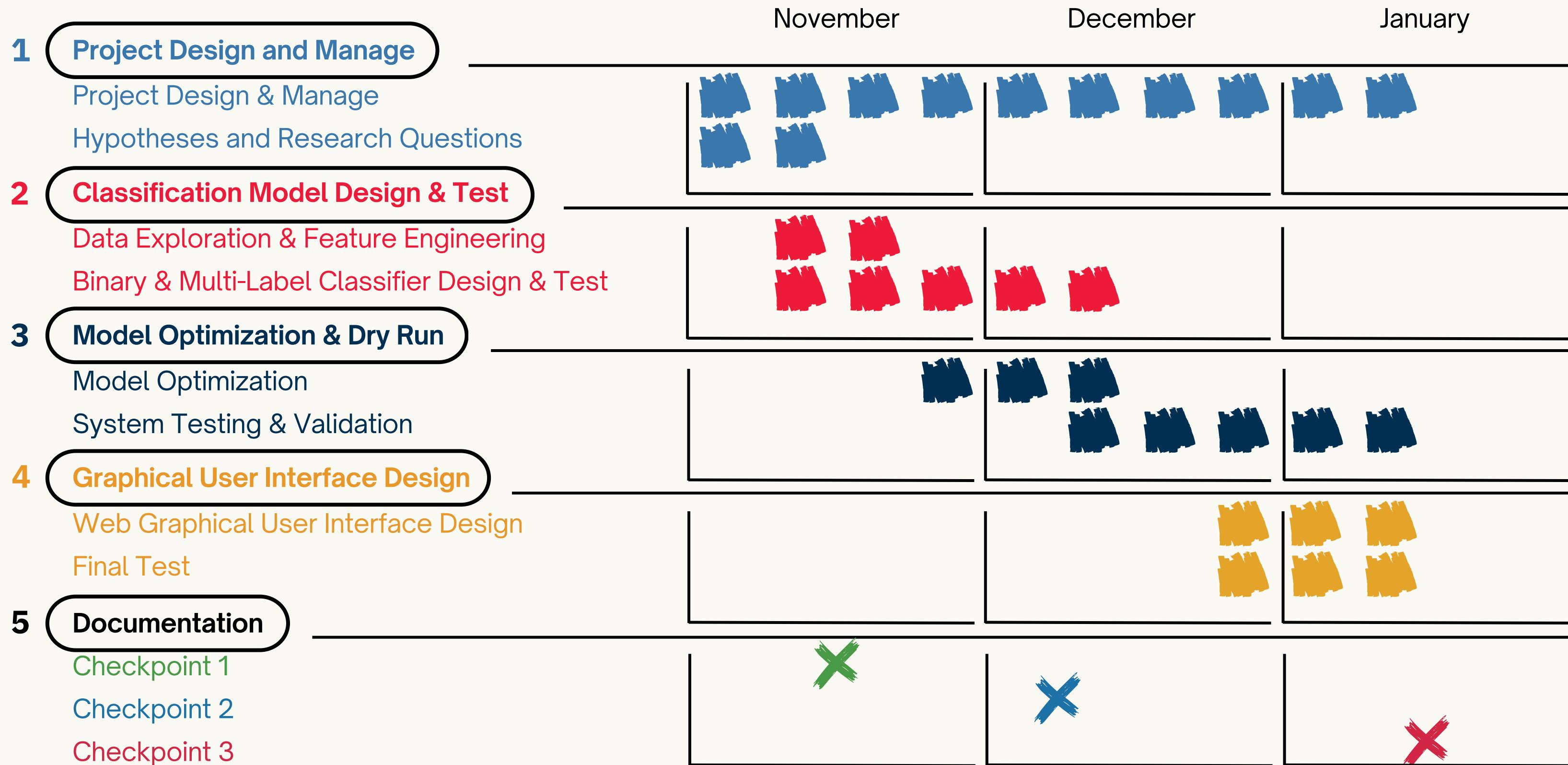


Table of Contents

Part 2

Project
Picture

Project
Objectives

Binary
Text Classification Problem

Multi-Label
Text Classification Problem

Project
Pipeline Design

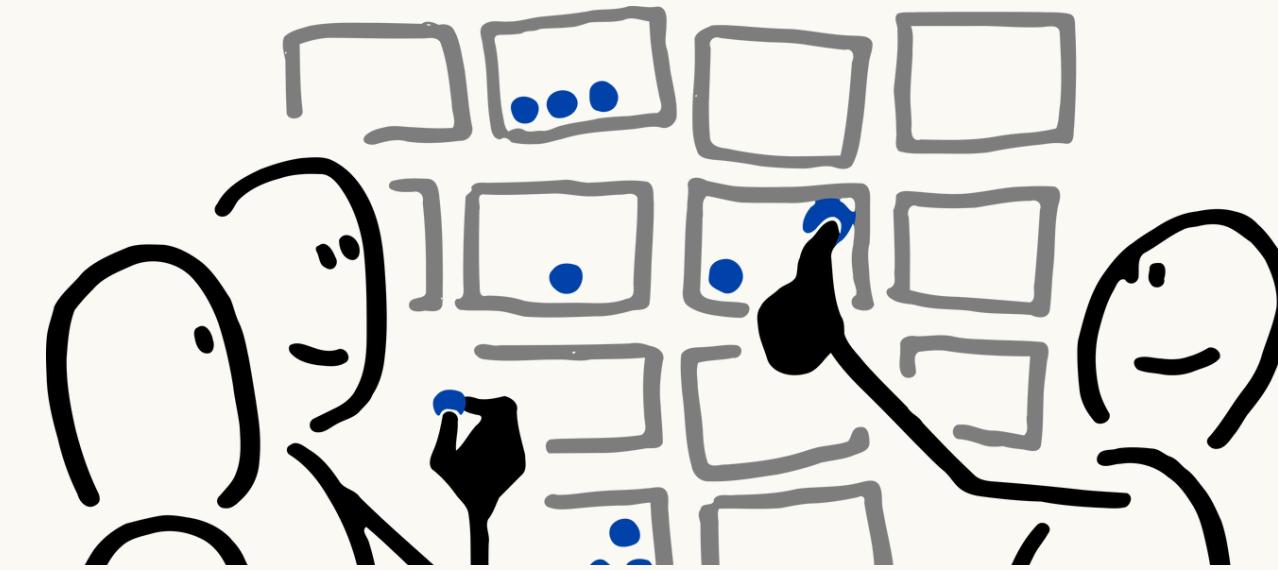
The **Picture**

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

- EU Commission



The **Project Objective**



Classifier Model for Scientific Papers



Detect Patient Preference Studies from PubMed outputs

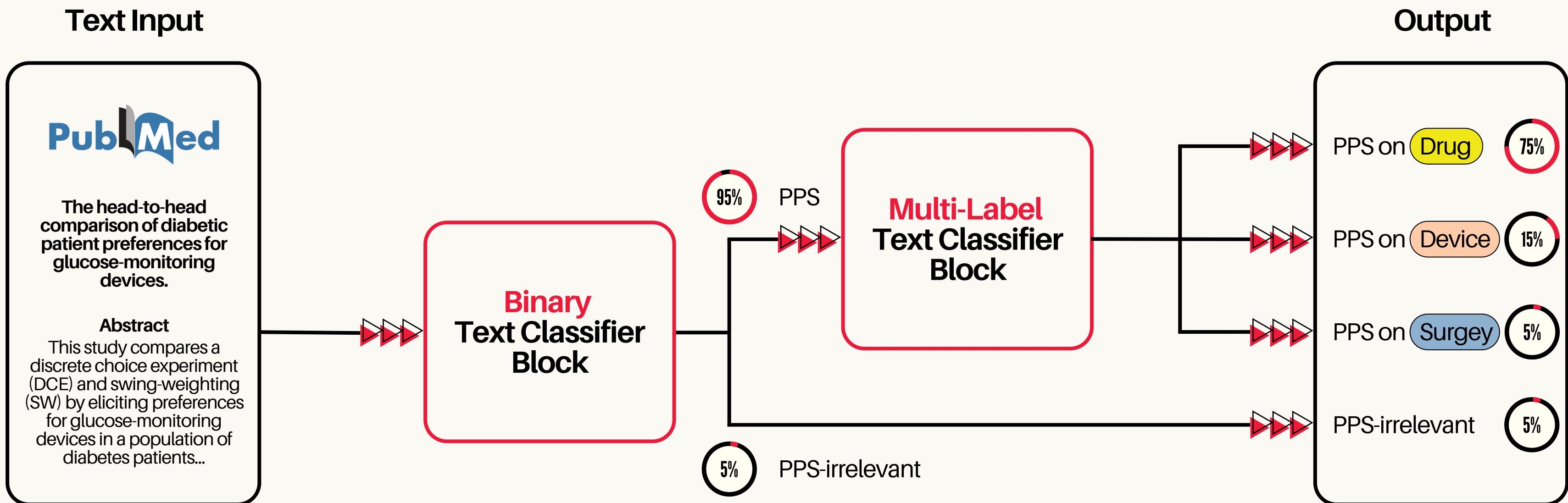


Categorize PPS-relevant papers into **Clinical Areas**

by **Titles** and **Abstracts**

The Task

Two-Stage Text Classification

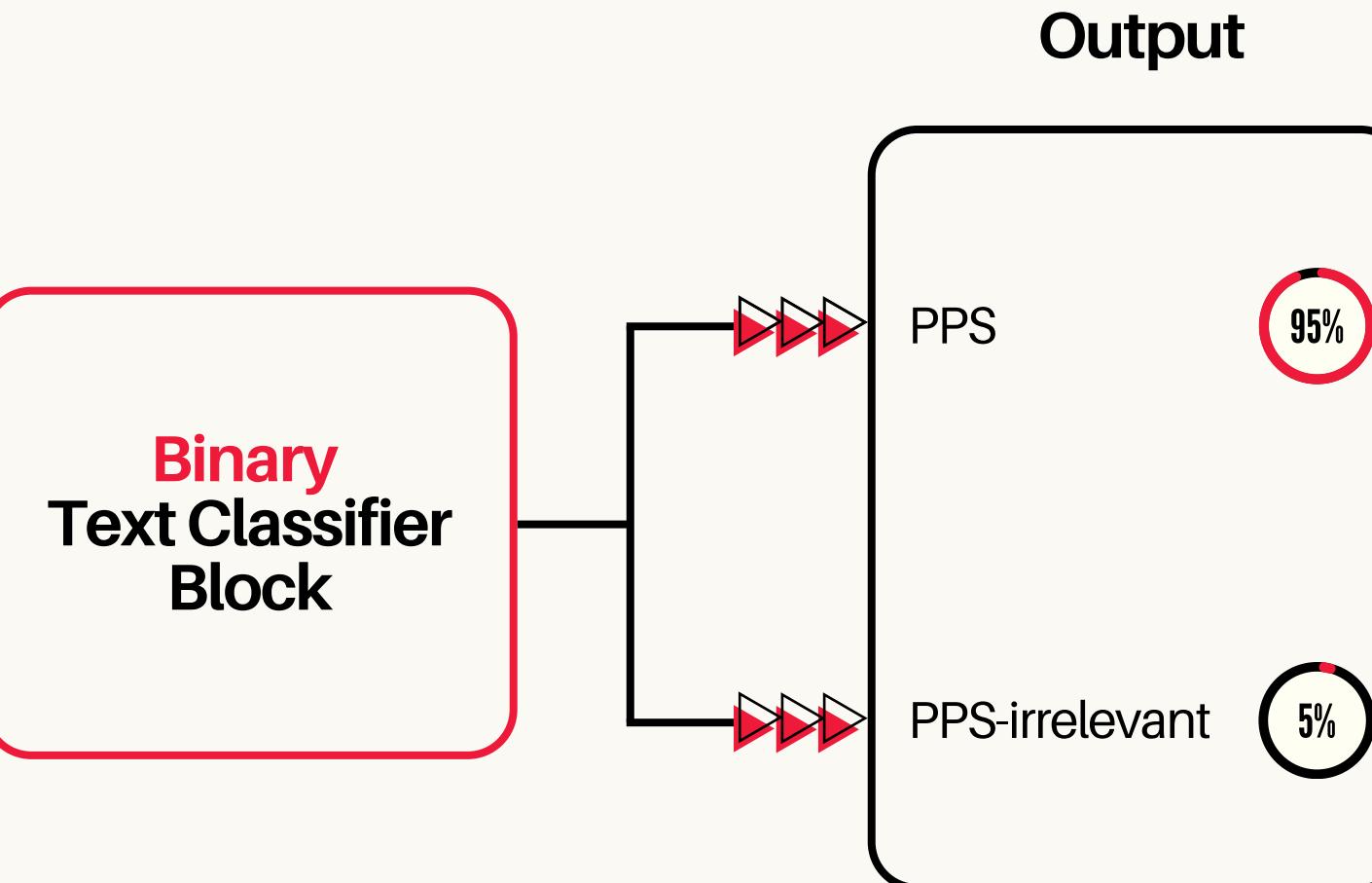
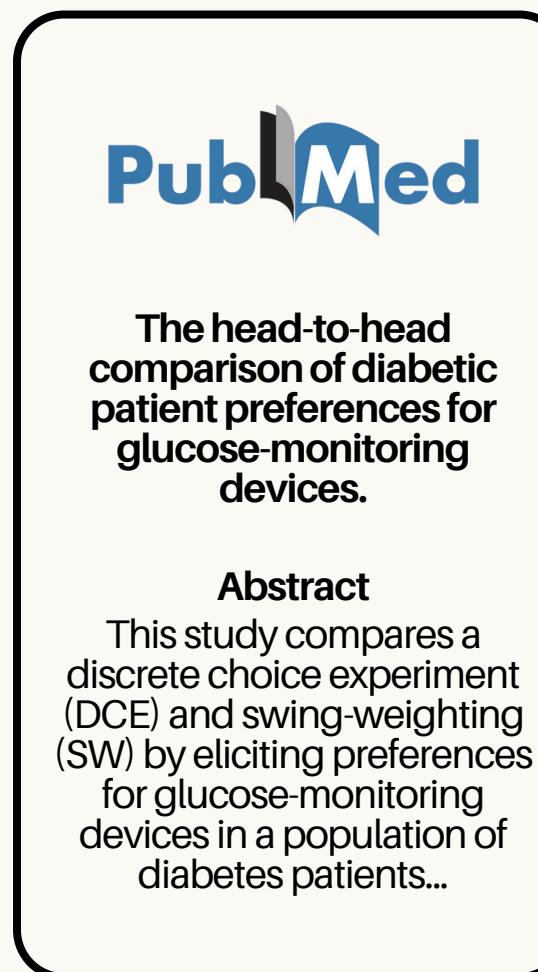


The Binary Text Classifier

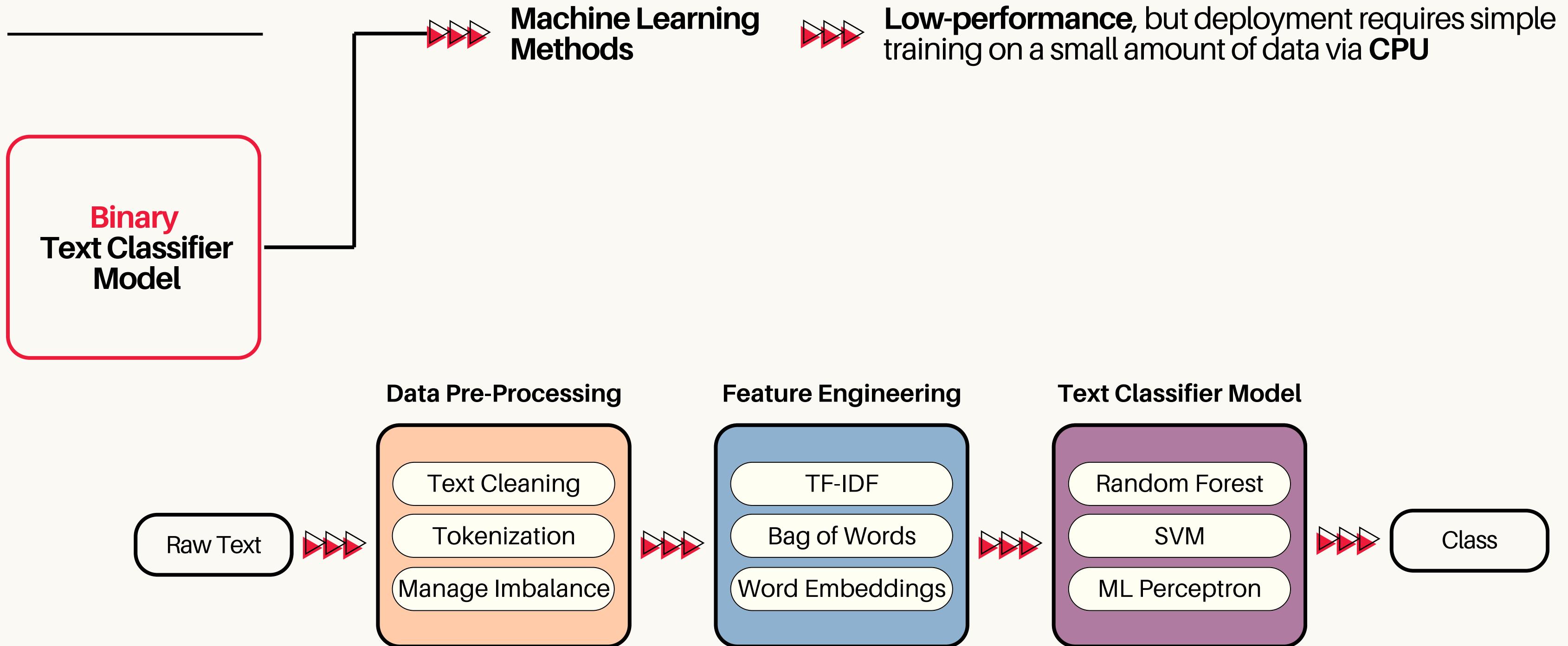


Detect Patient Preference Studies from PubMed outputs
by Titles and Abstracts

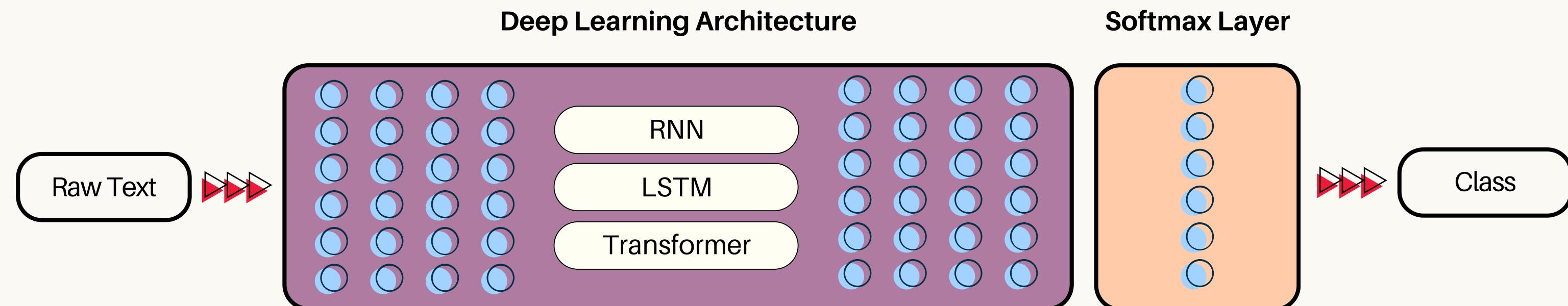
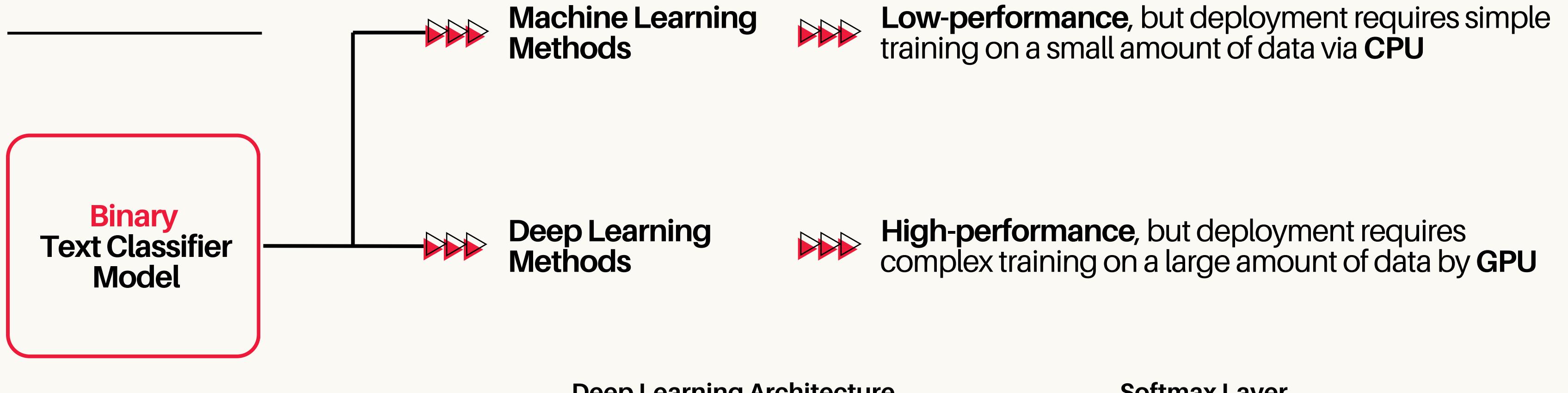
Text Input



The Binary Classifier Methods

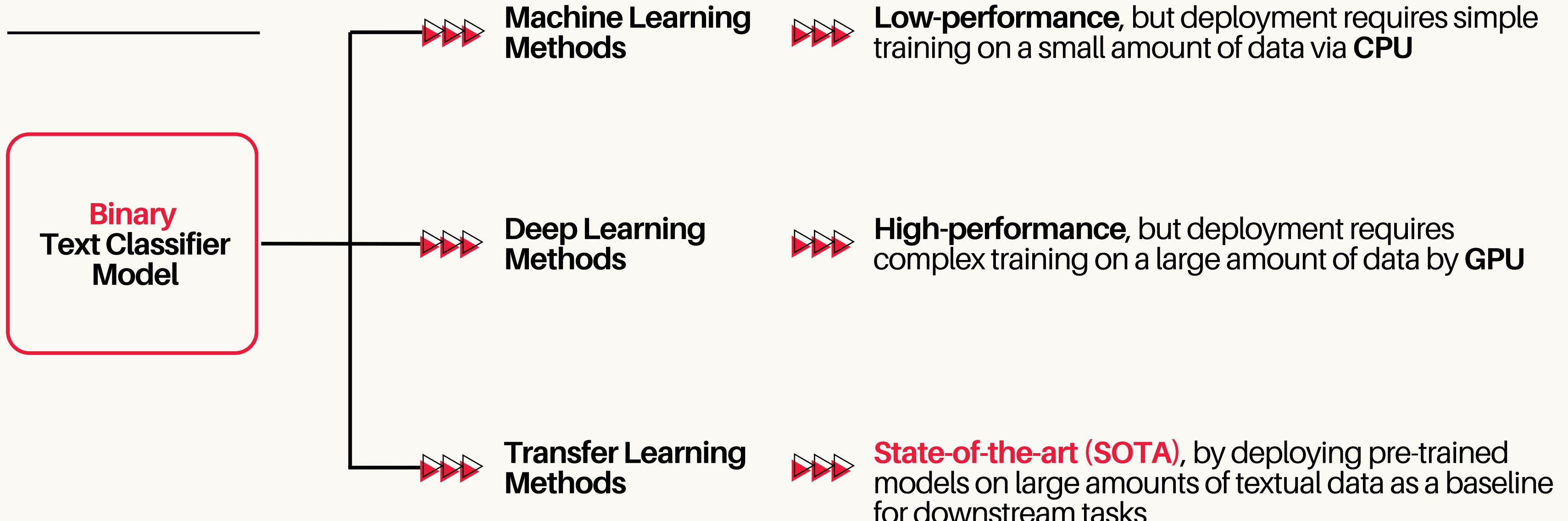


The Binary Classifier Methods

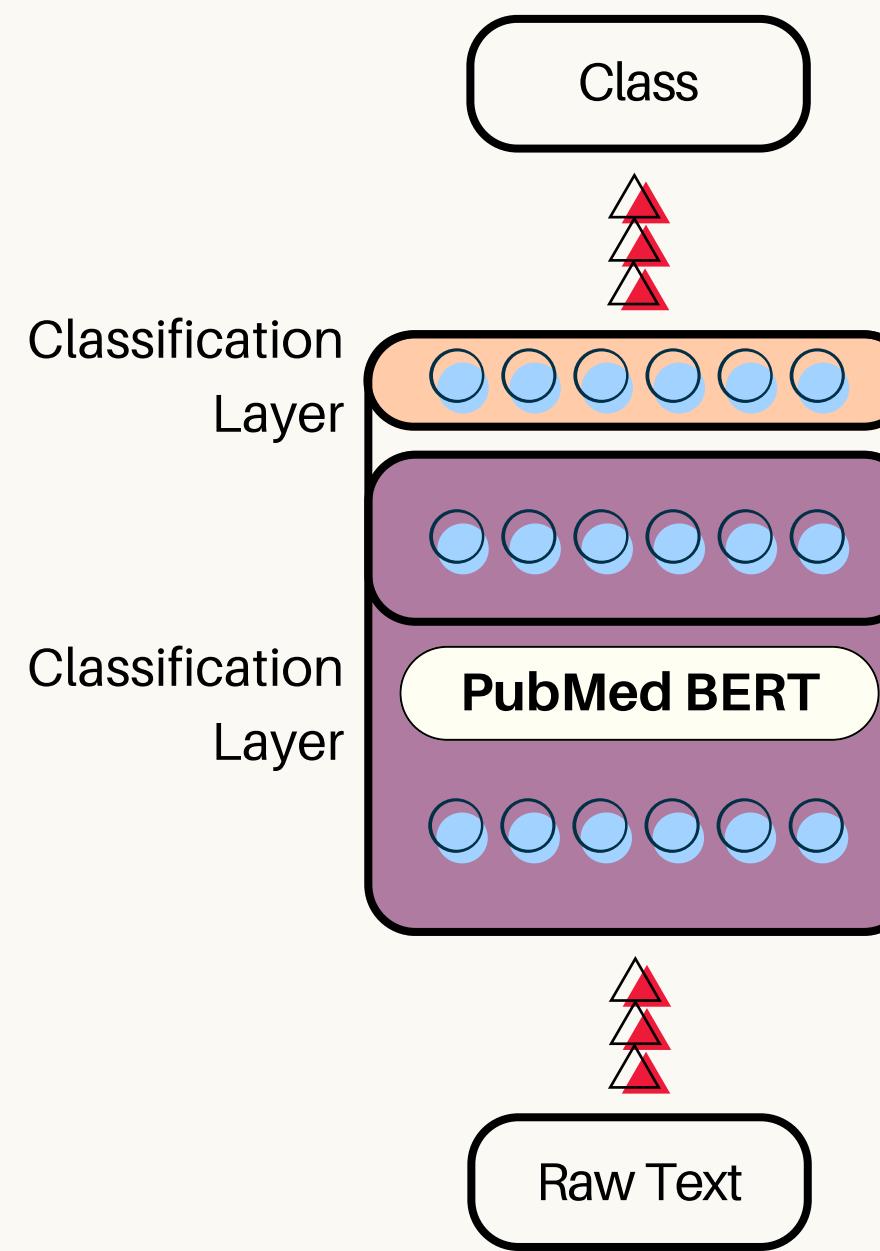


Binary Text Classification Problem

The Binary Classifier Methods

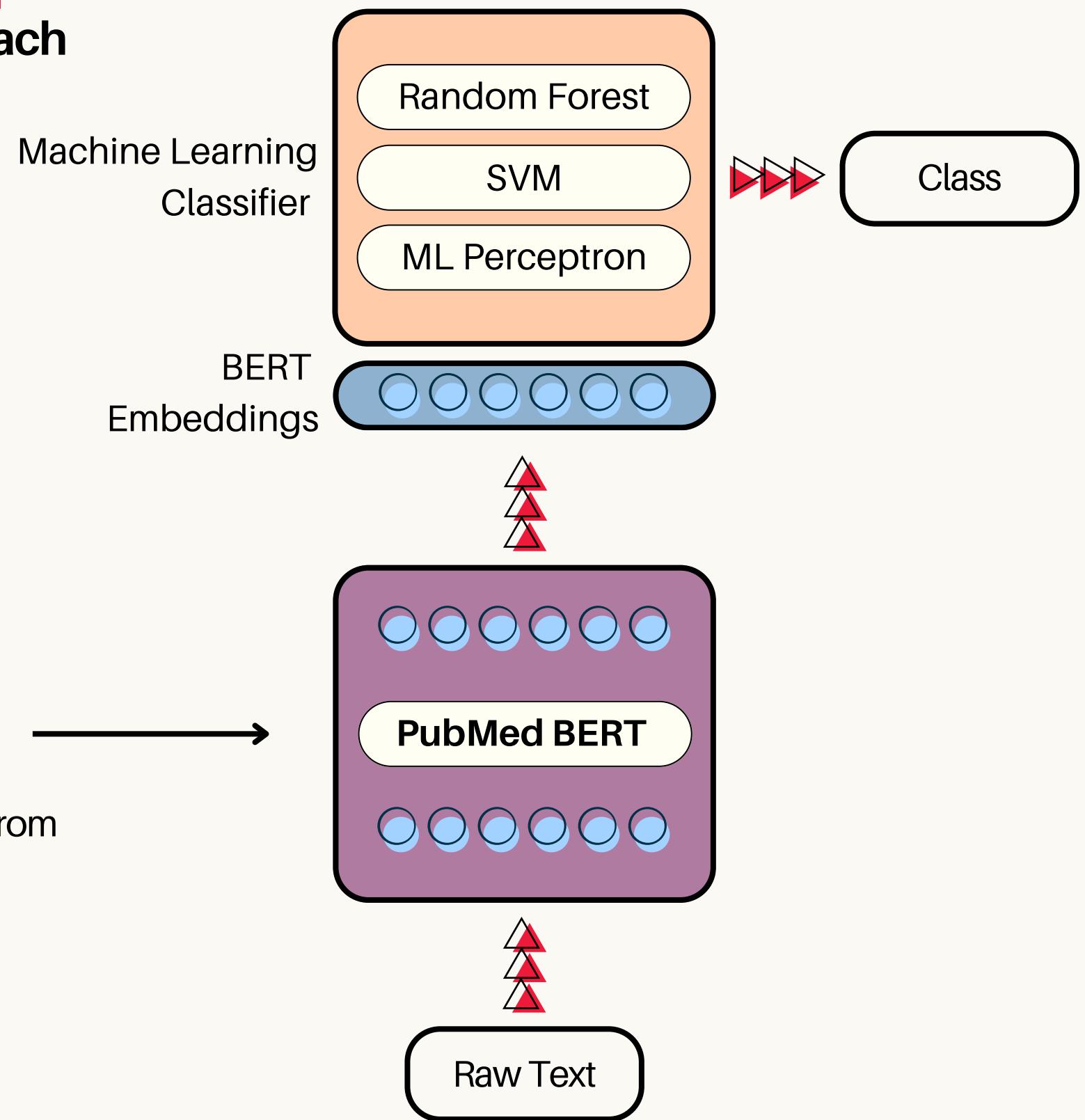


The Choices



The Standard Approach

The Hybrid Approach



Pre-Trained BERT Model

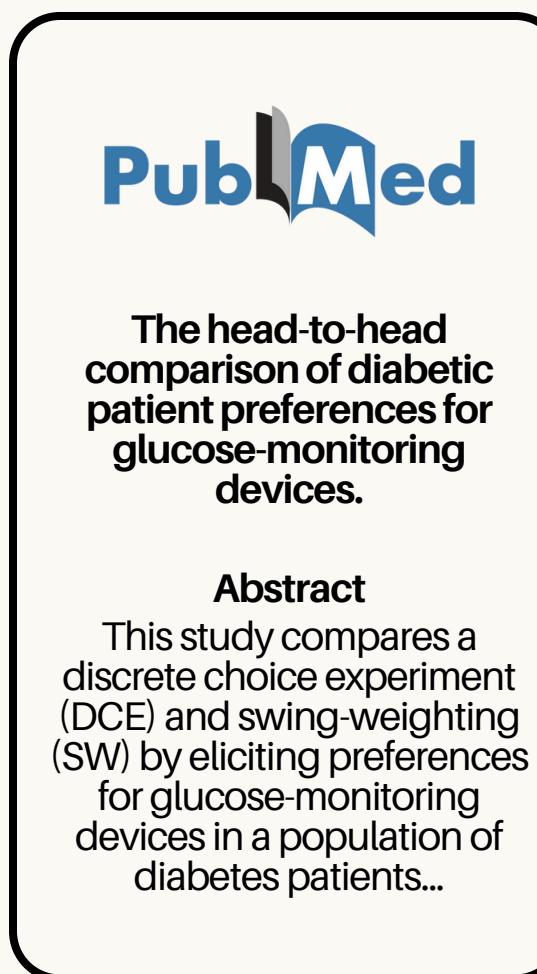
from scratch on abstract from
PubMed

The Multi-Label Text Classifier

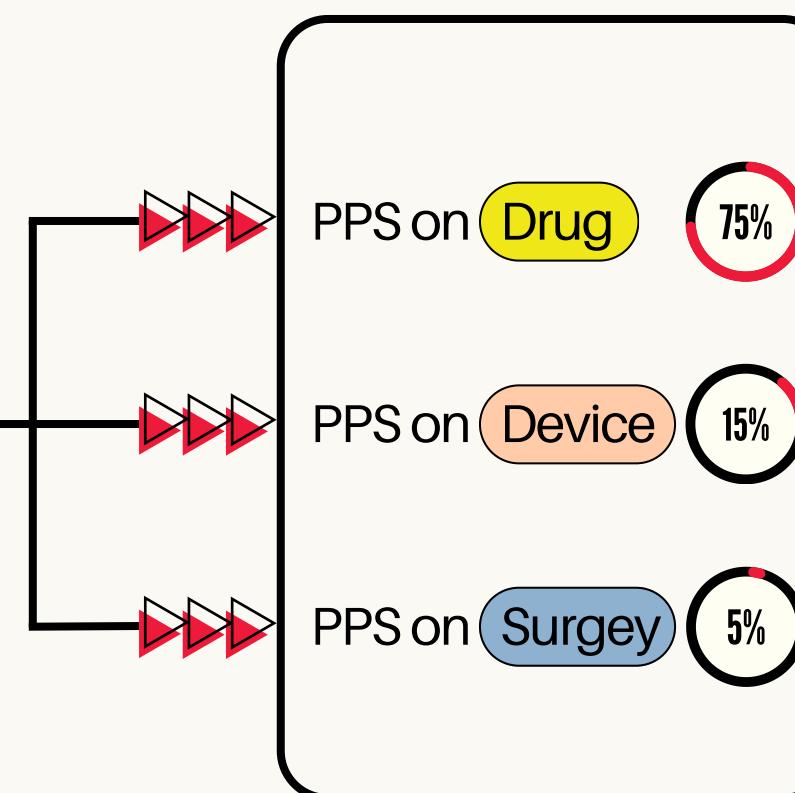


Categorize PPS-relevant papers into **Clinical Areas**
by **Titles and Abstracts**

Text Input



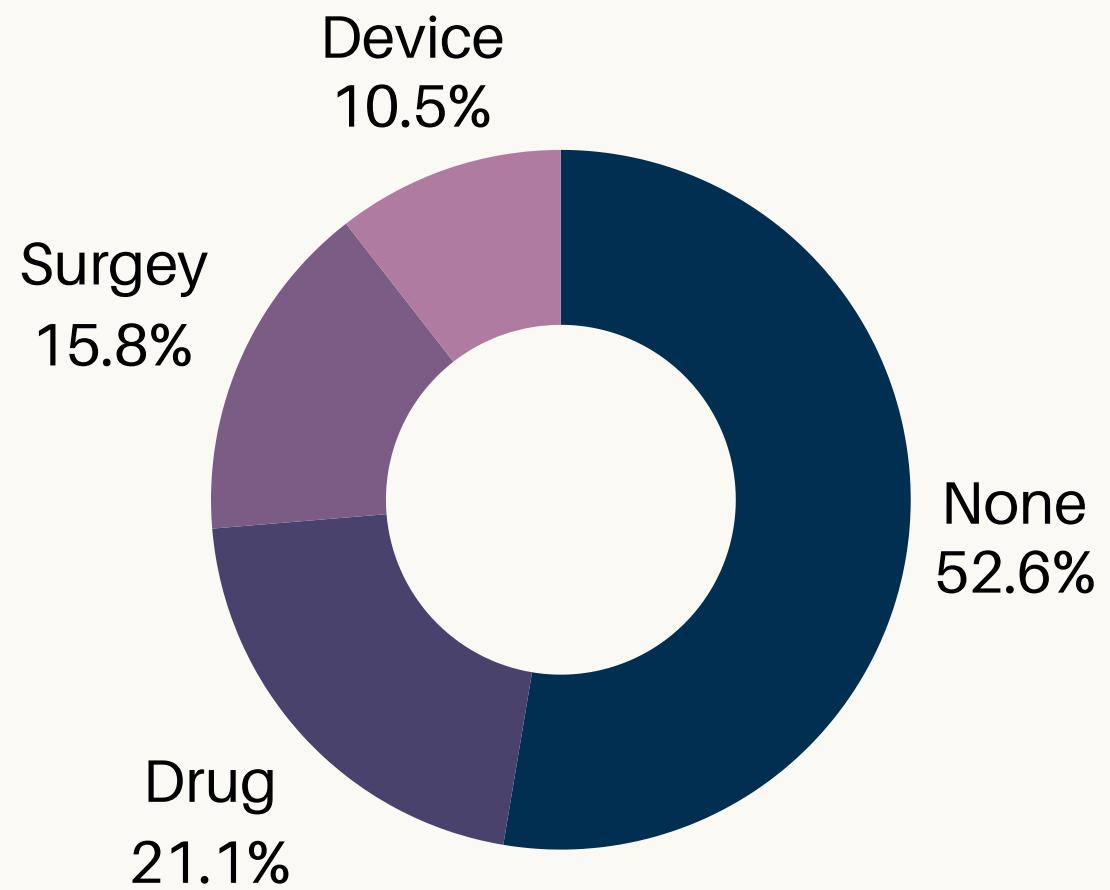
Output



**Multi-Label
Text Classifier
Block**

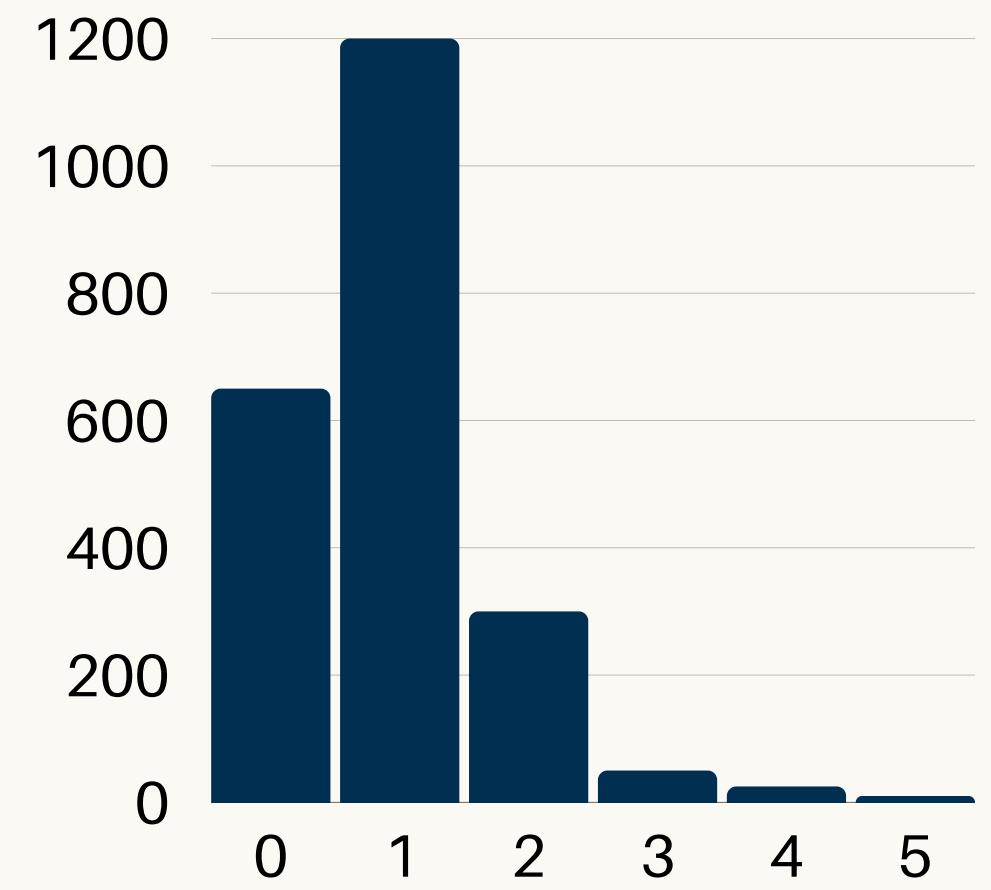
The Multi-Label Classifier Problem

Class Distribution



Data Imbalance

Multi-Label Distribution



Research **Questions**

“What are the most effective techniques for **building a multi-label classification model** to categorize scientific articles while addressing data imbalance?”

“How **can topic modeling** be used to determine if the identified **areas are sufficient, discover new areas** of interest and can the same model be leveraged for **classification?**”

The Multi-Labels Classifier Models

Problem Adaptation

Algorithm Adaptation

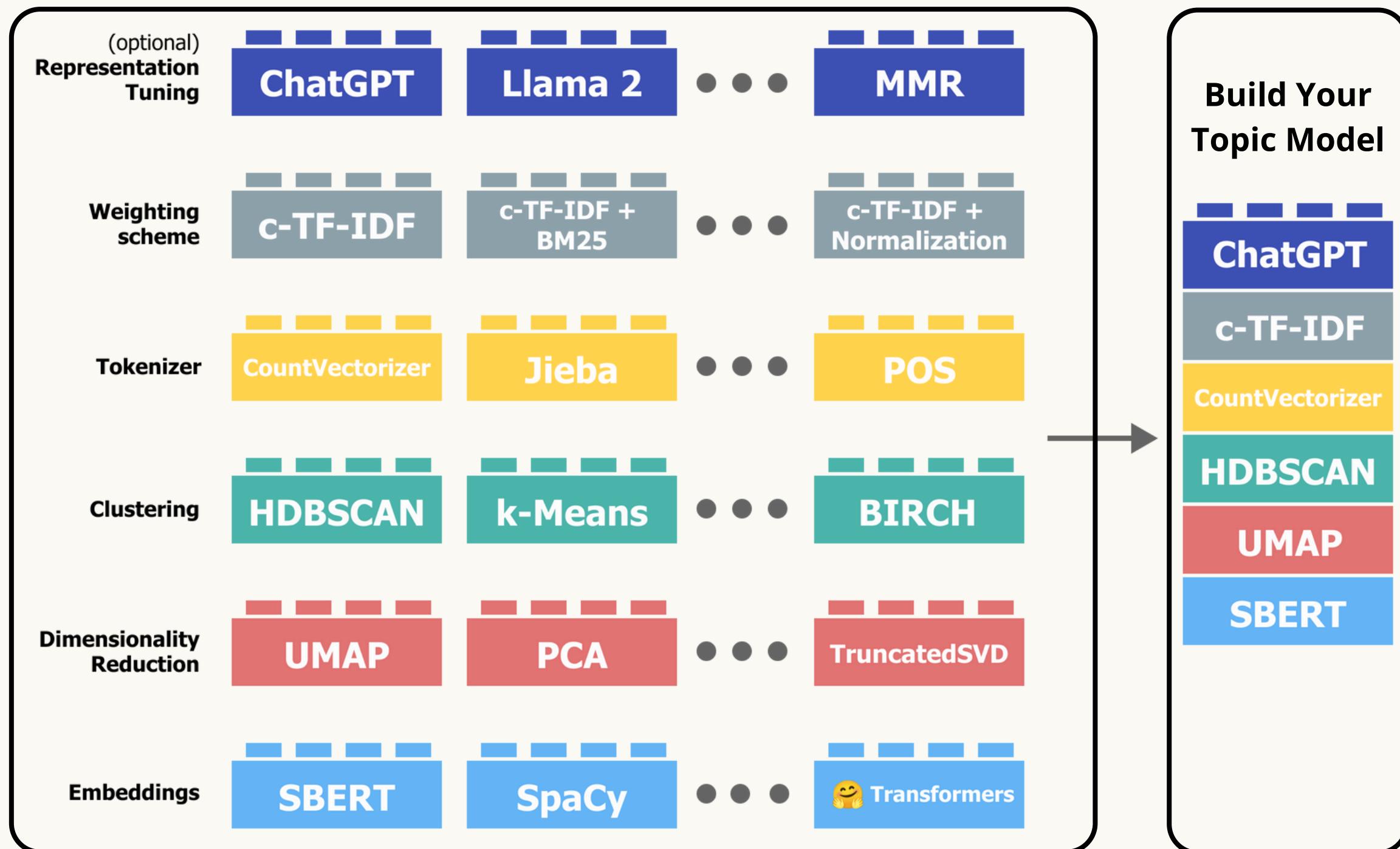
- Multi-Label k-NN
- Multi-Class MLP
- Ranking SVM

Topic Modelling

- LDA
- NMF
- Top2Vec
- **Bertopic**



Multi-Label
Text Classification Problem



BERTopic Project Benefits

▶▶▶ Topic Exploration

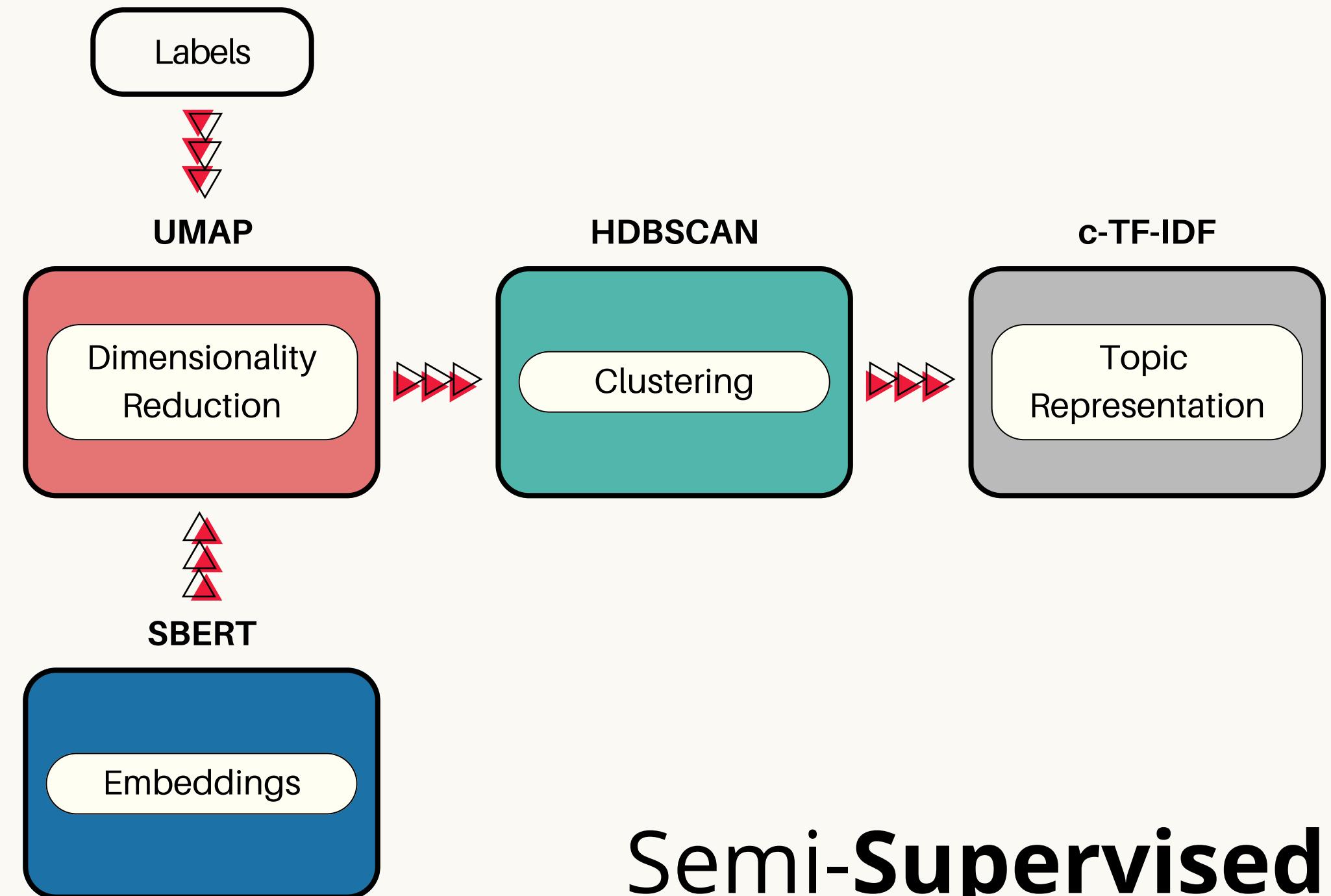
▶▶▶ Categories Not Fixed

▶▶▶ Improve & Speed Up
Manual Labeling

- Binary Classifier
- Multi-Label Classifier

▶▶▶ Classification Approaches:

- Unsupervised
- Semi-Supervised
- Supervised: Classification



BERTopic Project Benefits

▶▶▶ Topic Exploration

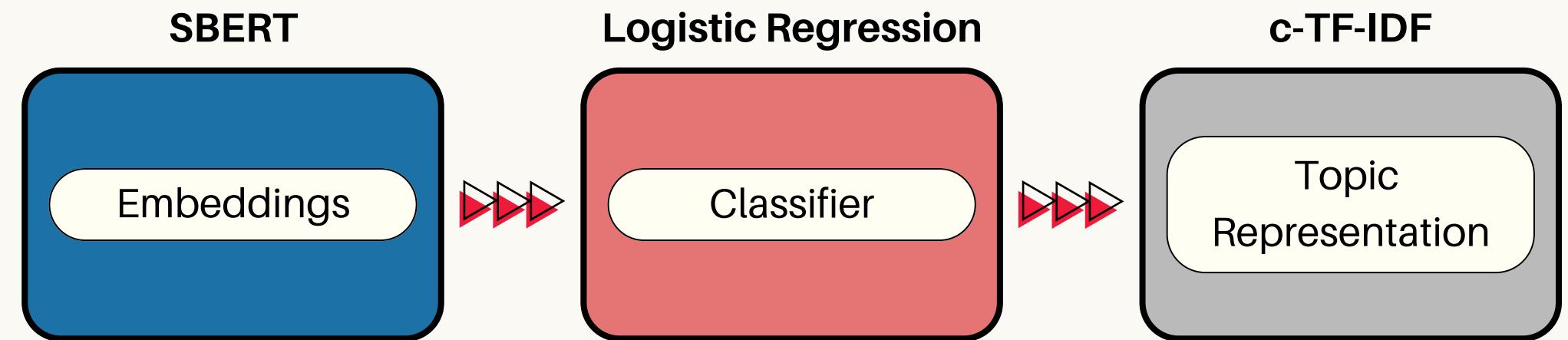
▶▶▶ Categories Not Fixed

▶▶▶ Improve & Speed Up
Manual Labeling

- Binary Classifier
- Multi-Label Classifier

▶▶▶ Classification Approaches:

- Unsupervised
- Semi-Supervised
- Supervised: Classification

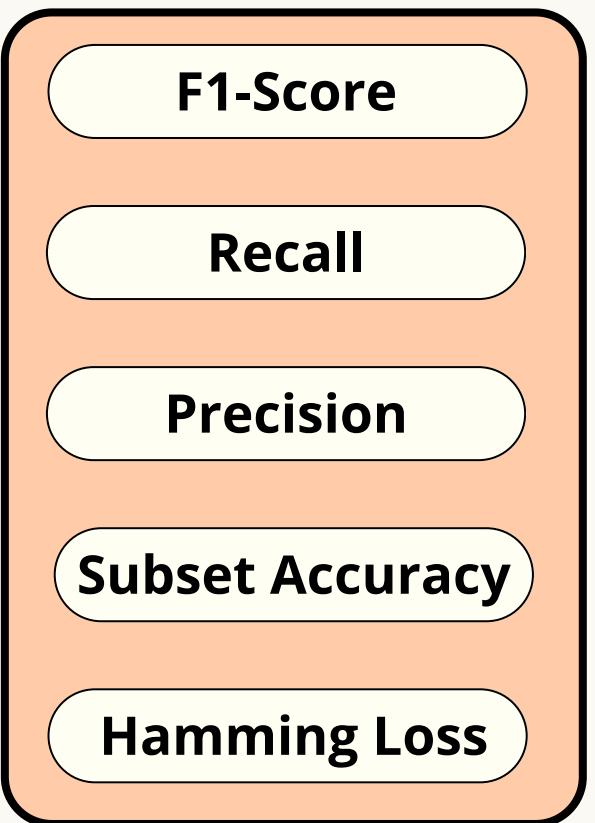


Supervised

The Evaluation Metrics

Classification Models

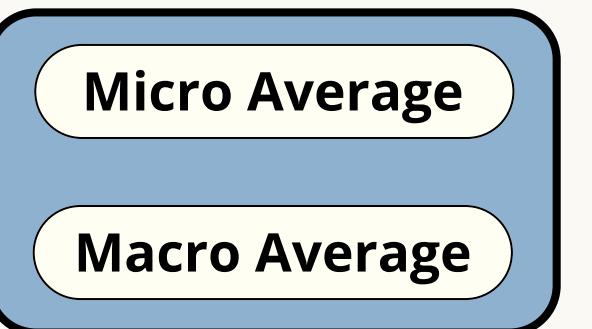
Example-Based



Ranking-Based



Label-Based



Topic Modelling

Diversity
Coherence

System Time

Train Time
Inference Time
Latency
Throughput
Scalability

The Design Pipeline

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

Abstract
This study compares a discrete choice experiment (DCE) and swing-weighting (SW) by eliciting preferences for glucose-monitoring devices in a population of diabetes patients...

Clinical Pubmed Article (PPS)

Data Augmentation

Data Pre-Processing

Text Cleaning
Handle Imbalance

Feature Engineering

PubMed BERT

Outliers
Label-Free Data

Model Optimization

Params Tuning
Cross Validation

Evaluation

Label-Based
Example-Based
Ranking-Based
Time-Based

Output

Predicted Multi-Label Classifications

Help Data Pruning

Discover New Topics of Interest

Modeling MLP Classifiers

Multi-Label k-NN
Multi-Label MLP
Ranking SVM
BERTopic
Topic Modeling

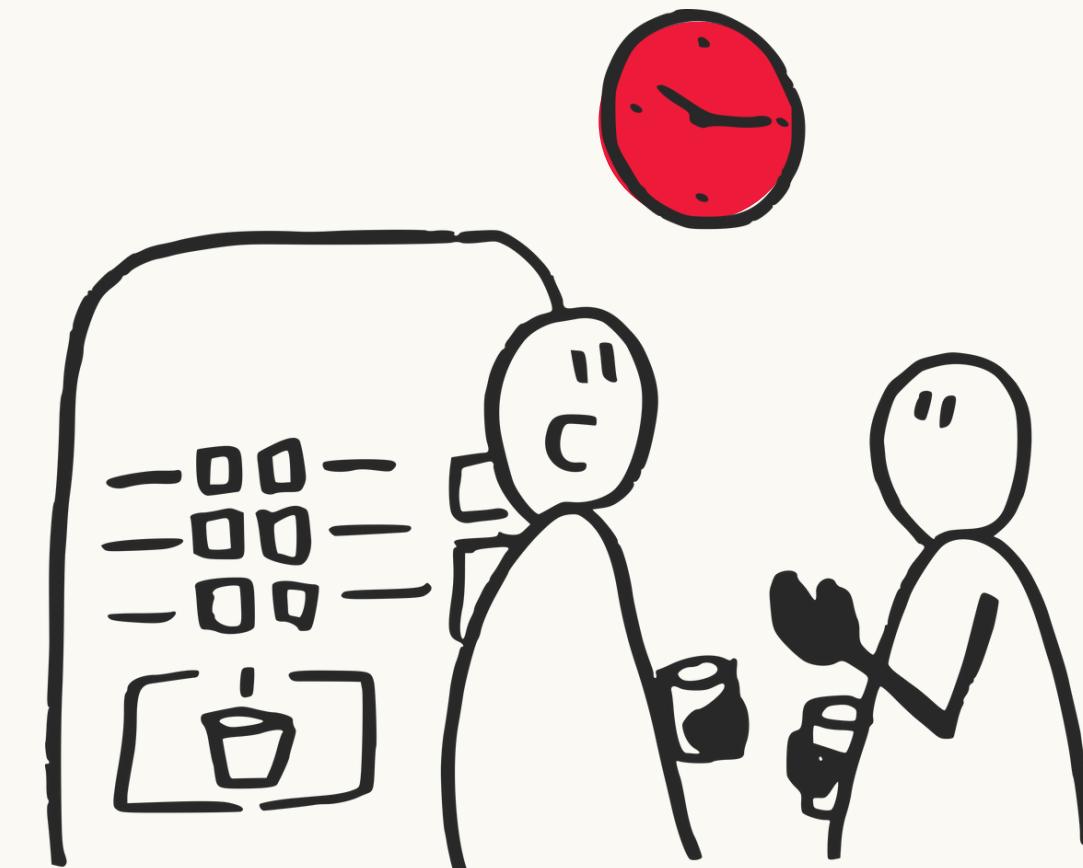
APPLIED DATA SCIENCE PROJECT

Thank You

Francesco Giuseppe Gillio & Cesar Augusto Seminario Yrigoyen



UNIVERSITÀ
DI TORINO



Politecnico
di Torino