# Classification and Retrieval System for Gas Pipe Repairs

*CheckPoint #2*

*Team 1: Michele Panteleo, Minal Jamshed, Prima Acharjee*

FONDAZIONE links
PASSION FOR INNOVATION

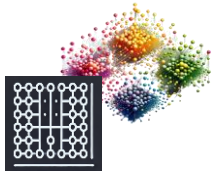
CoRe

1859

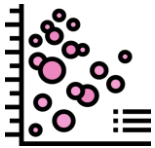
# Outline

Recap

Data Preparation



Methodology



Evaluation

What's next?

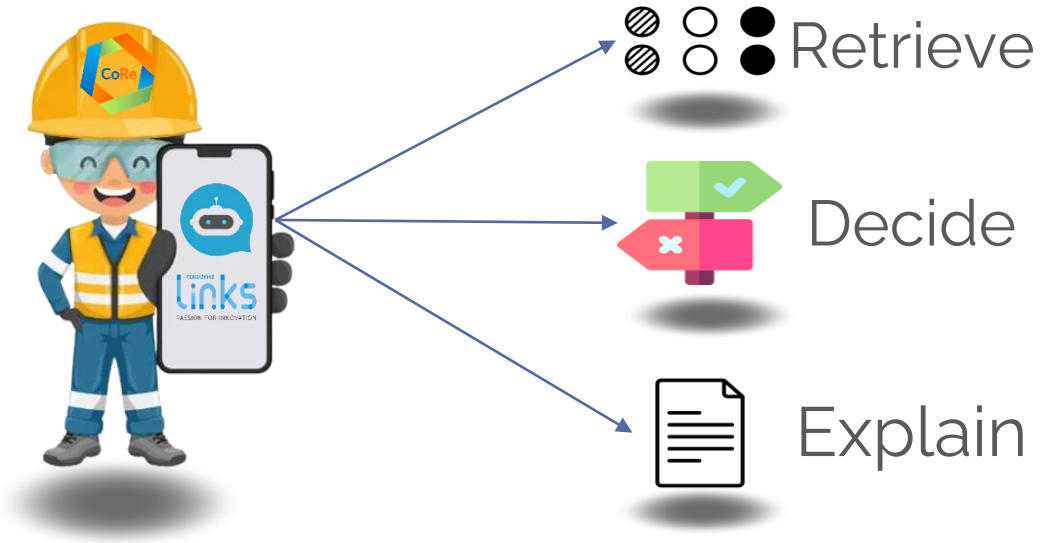
# Problem Description

This project deals with **gas leaks.** Whenever a fault happens, gasfitters are asked to **fix** it by choosing the **best** strategy (welding, substituting, **patching**).

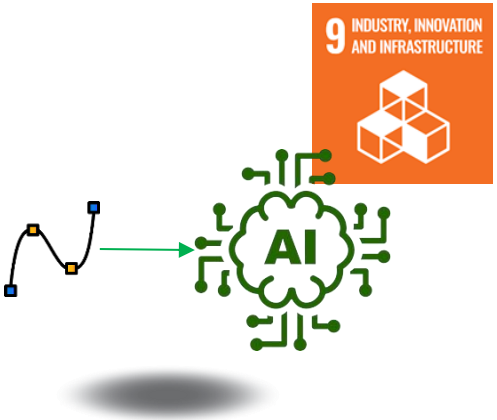
Gas leak

# Mission

Our mission is to develop a **chatbot** to help gasfitters about **patching or not,** explaining **why**. Chatbot will answer based on **fault description** and **past interventions**.

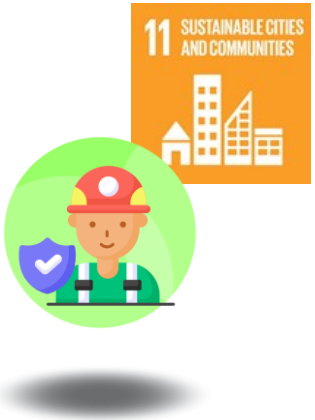


# Value Proposition




Innovate


Reduce Gas Leaks Impact

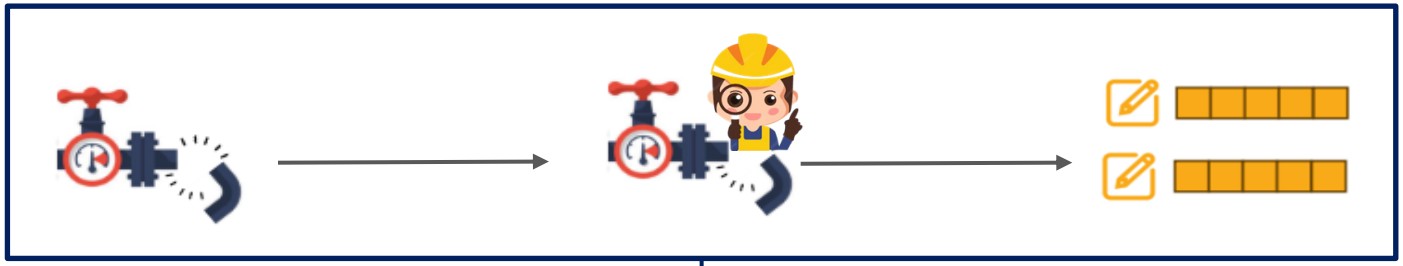



Improve Safety

# Data Preparation

# Data Collection

# Data Collection



| ID | Repair_ Code | Kit_Size _num | Label | Kit_Compliance | Damage_Type | Pipe_Exposure | Severe_Corrosion | Pipe_Covered | Damaged_Valve | Summary | Summary_1 | Kit_Size_str |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID000001 | INT000001 | 5 | FALSE | The size and ty | Non-sheared lin | Aerial pipe | TRUE | TRUE | TRUE | The aerial | The aerial stee | XL |
| ID000002 | INT000002 | 5 | FALSE | The size and ty | Non-sheared lin | Aerial pipe | TRUE | TRUE | TRUE | A non-shea | A localized line | XL |
| ID000003 | INT000003 | 5 | FALSE | The size and ty | Non-sheared lin | Aerial pipe | TRUE | TRUE | FALSE | Aerial stee | Aerial steel pi | XL |
| ID000004 | INT000004 | 5 | FALSE | The size and ty | Non-sheared lin | Aerial pipe | TRUE | TRUE | FALSE | The aerial | The steel aeria | XL |

Collect and
Organize in a CSV

# Data Collection

| ID | Repair_ Code | Kit_Size _num | Label | Kit_Compliance | Damage_Type | Pipe_Exposure | Severe_Corrosion | Pipe_Covered | Damaged_Valve | Summary | Summary_1 | Kit_Size_ str |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID000001 | INT000001 | 5 | FALSE | The size and ty | Non-sheared lin | Aerial pipe | TRUE | TRUE | TRUE | The aerial | The aerial stee | XL |
| ID000002 | INT000002 | 5 | FALSE | The size and ty | Non-sheared lin | Aerial pipe | TRUE | TRUE | TRUE | A non-shea | A localized line | XL |
| ID000003 | INT000003 | 5 | FALSE | The size and ty | Non-sheared lin | Aerial pipe | TRUE | TRUE | FALSE | Aerial stee | Aerial steel pip | XL |
| ID000004 | INT000004 | 5 | FALSE | The size and ty | Non-sheared lin | Aerial pipe | TRUE | TRUE | FALSE | The aerial | The steel aeria | XL |

Features considered

# Two Modalities

**Summary**

False: 11.778 (99%)
True: 126 (1%)

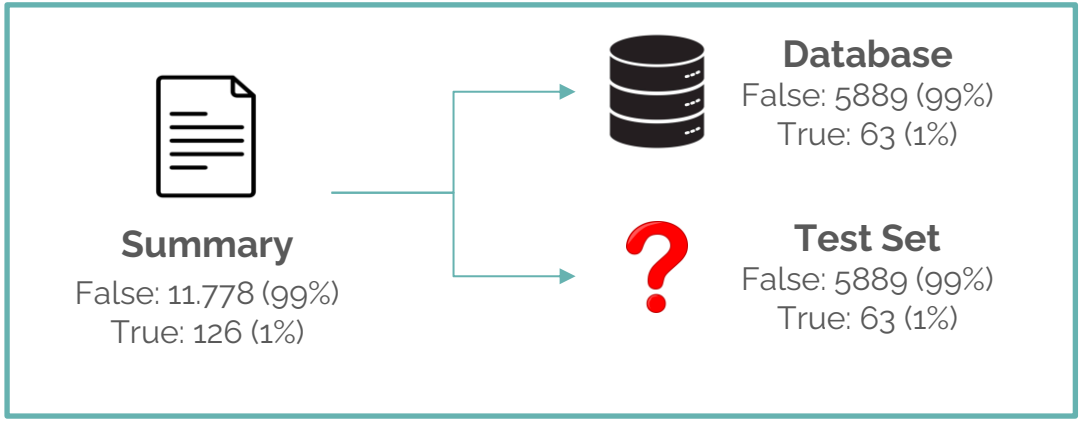
# Two Modalities

**Summary**
False: 11.778 (99%)
True: 126 (1%)

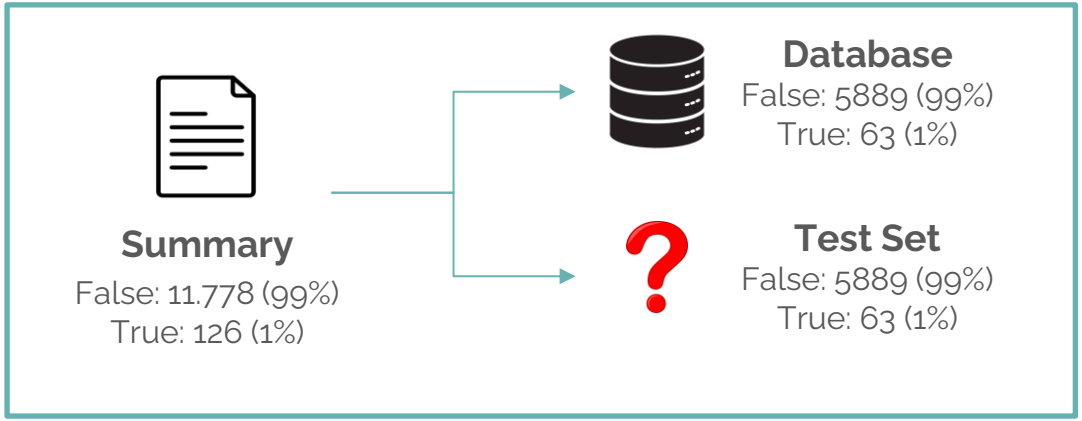
**Database**
False: 5889 (99%)
True: 63 (1%)

**Test Set**
False: 5889 (99%)
True: 63 (1%)

# Two Modalities




## Balanced Split

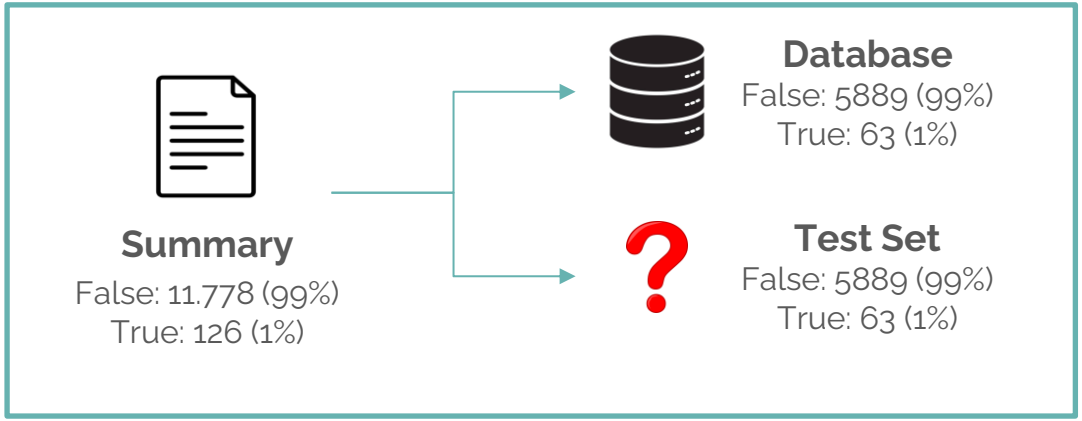



**Balanced Split**

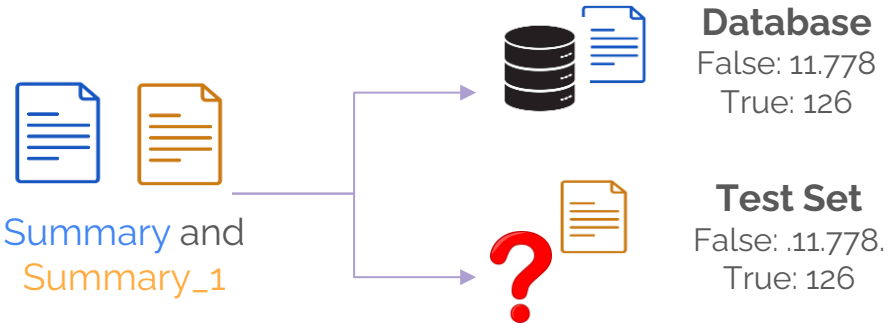
**Two Modalities**

Summary and Summary_1

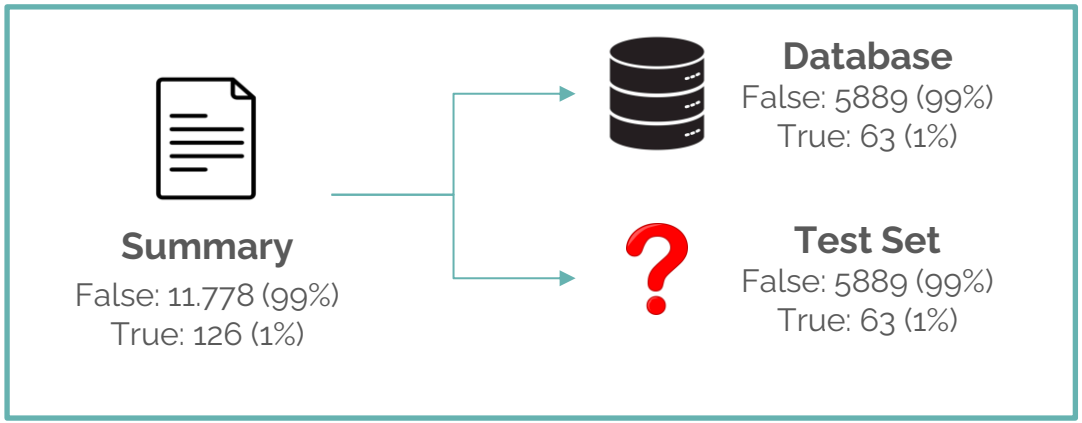
# Two Modalities



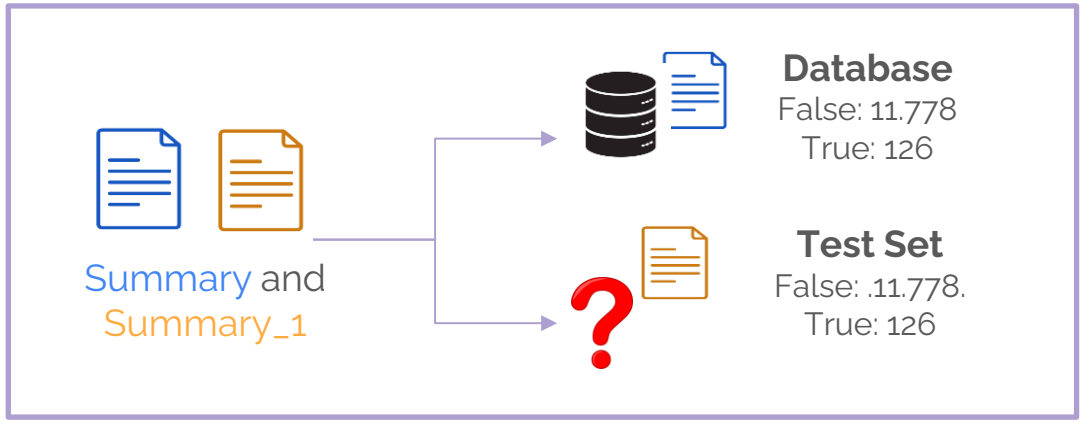

**Balanced Split**



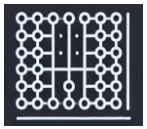
# Two Modalities




## Balanced Split

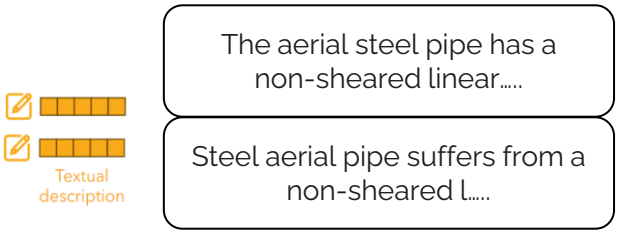



## Dedicated Split



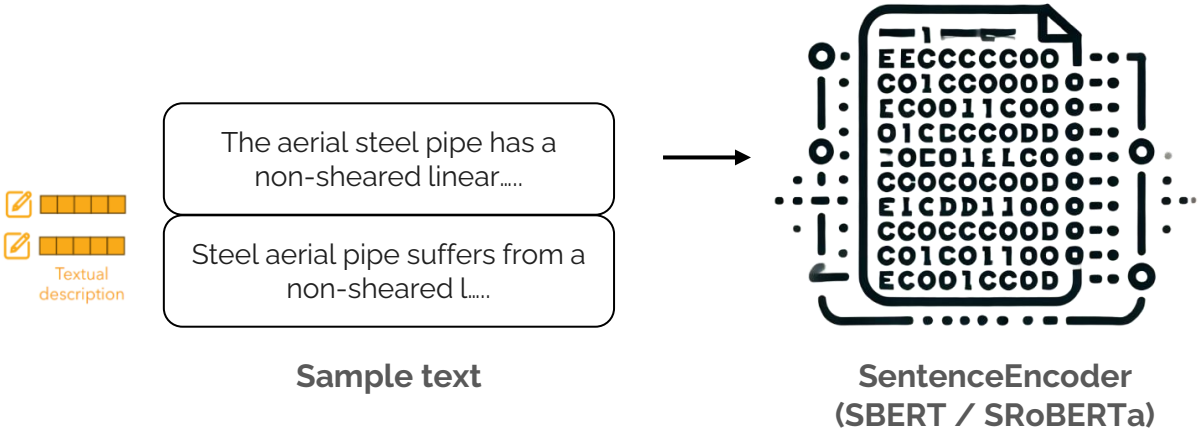
# Methodology

# Generating Embeddings with SBERT/SRoBERTa

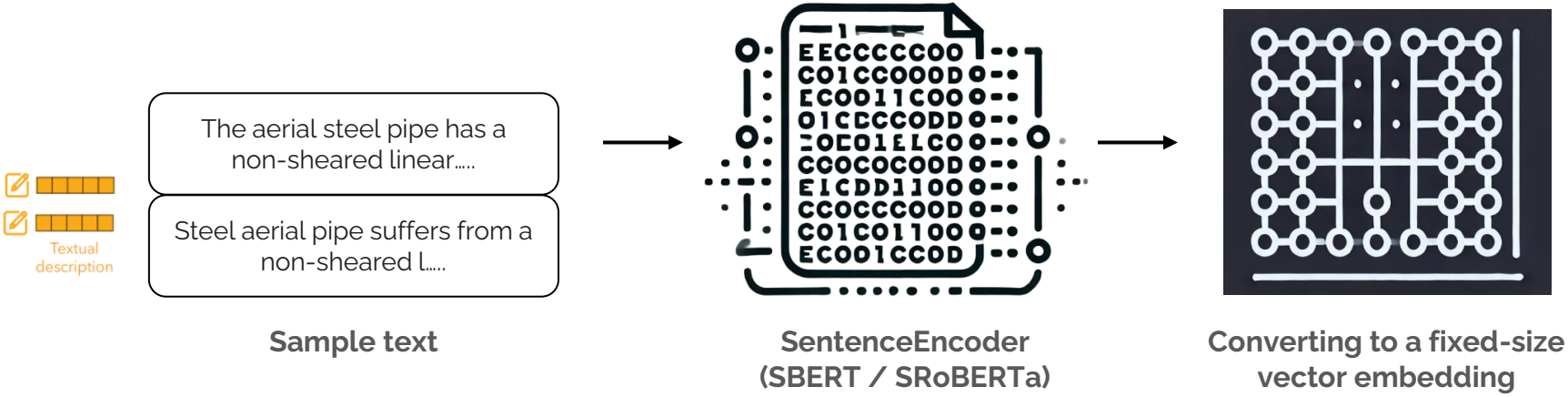



**Sample text**

# Generating Embeddings with SBERT/SRoBERTa



# Generating Embeddings with SBERT/SRoBERTa

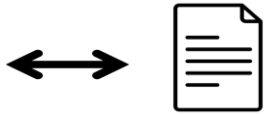


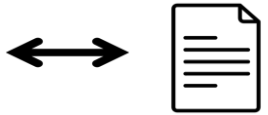
# Organizing Embeddings in DataFrames

Database

Test Query Set

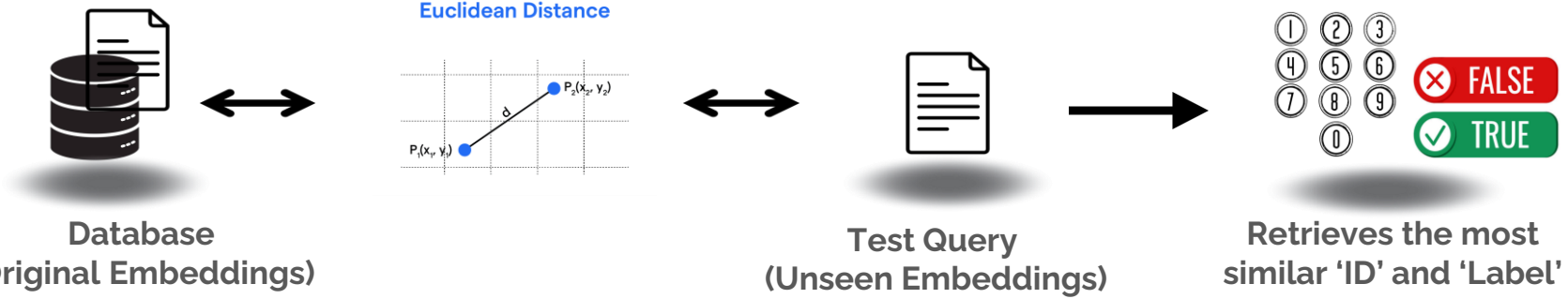
| ID | Embeddings | Successful |
|---|---|---|
| ID000064 | [-0.001362897455692913, 0.9544798135757446, 0.0...] | TRUE |
| ID000128 | [1.3155205249786377, -0.04855910316109657, 0.0...] | FALSE |
| ID000192 | [-1.1072819232490674, 0.05280644819140434, 0.18...] | FALSE |
| ID000453 | [1.0781232490879674, 0.5228344819140434, 0.23...] | TRUE |
| ID007895 | [-0.023719232490674, 0.01652804819140434, 0.39...] | TRUE |





Save as CSV Files

# Similarity Checking-FAISS

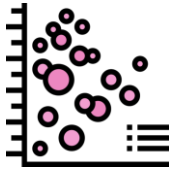



Database
(Original Embeddings)

Test Query
(Unseen Embeddings)

Retrieves the most
similar ‘ID’ and ‘Label’

# Evaluation



# SBERT vs. SRoBERTa Across Modalities

## Retrieval of IDs Performance of *Modality-2*: SBERT vs. SRoBERTa

| Model | Recall@1 | Recall@3 |
|---|---|---|
| **SBERT** | 0.61 | 0.69 |
| **SRoBERTa** | **0.77** | **0.84** |

**Recall@k computation=**

$$\frac{\text{Number of recommended relevant items among top k}}{\text{Number of all relevant items in the system}}$$

# Retrieval Task through SBERT vs SRoBERTa - Recall@1

**Comparison-** summarizing key metrics side-by-side

| Metric | SBERT | SRoBERTa |
|---|---|---|
| True Positive (TP) | 7291 | 9108 |
| False Negative (FN) | 4613 | 2796 |
| Recall@1 | **61%** | **77%** |

# SBERT vs. SRoBERTa Across Modalities

**Comparative metrics Table** for Classification of Labels

| Metric (in %) | M1 | | M2 | |
|---|---|---|---|---|
| | SBERT | SRoBERTa | SBERT | SRoBERTa |
| F1 score | 18 | 13 | 84 | 75 |
| Recall | 15 | 9.5 | 80 | 70 |
| Precision | 23 | 20 | 88 | 83 |
| Accuracy | 98 | 98 | 99 | 99 |

# How are the Metrics computed?

| Predicted Label \ Real Label | Positive | Negative |
|---|---|---|
| Positive | True Positive (TP) | False Positive (FP) |
| Negative | False Negative (FN) | True Negative (TN) |

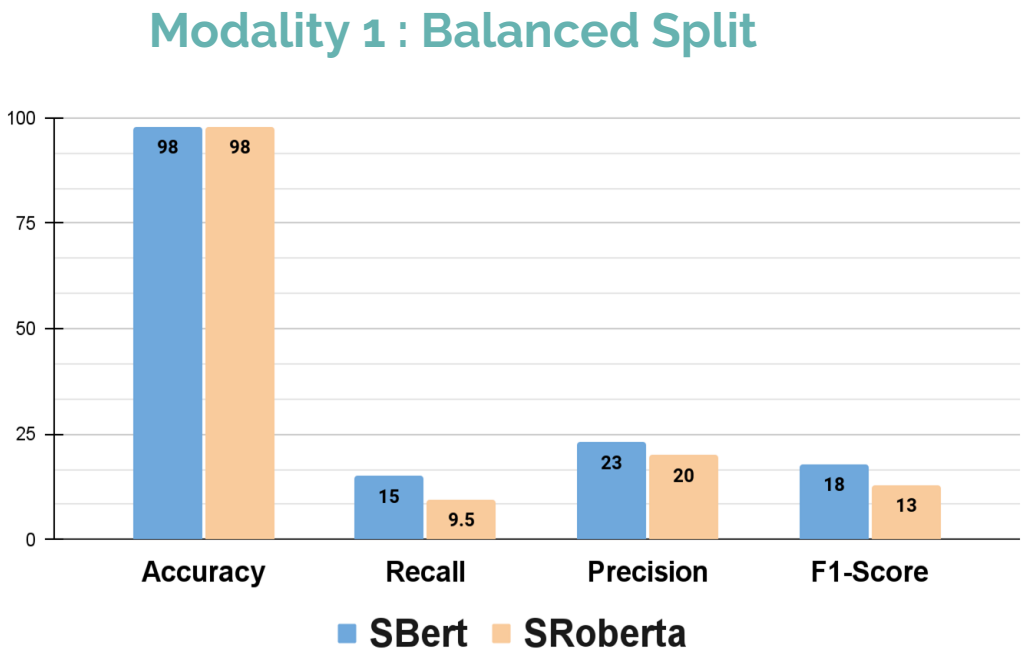
$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

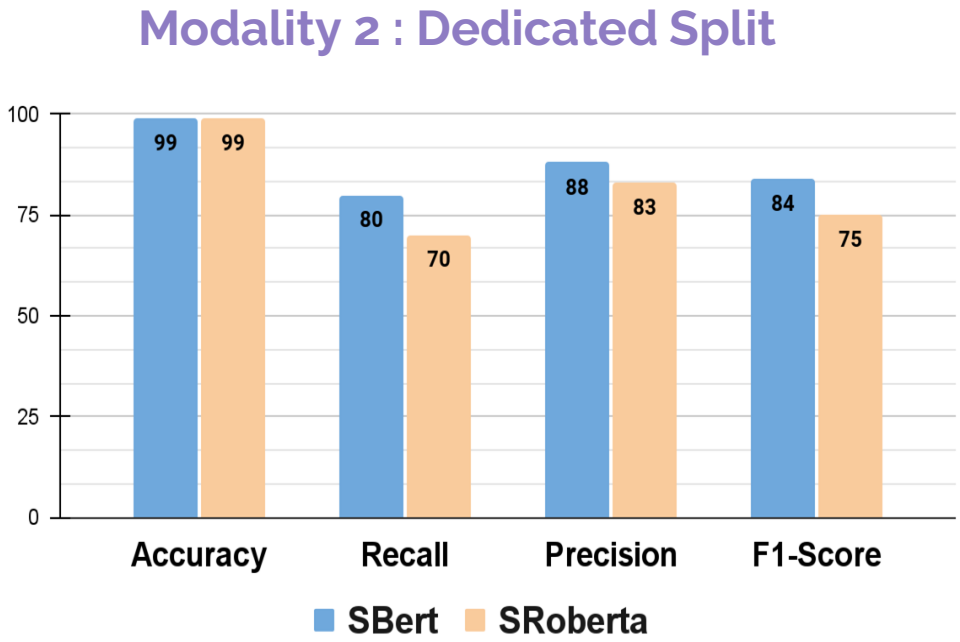
# SBERT vs. SRoBERTa Across Modalities

**Comparative metrics Graph** for Classification of Labels



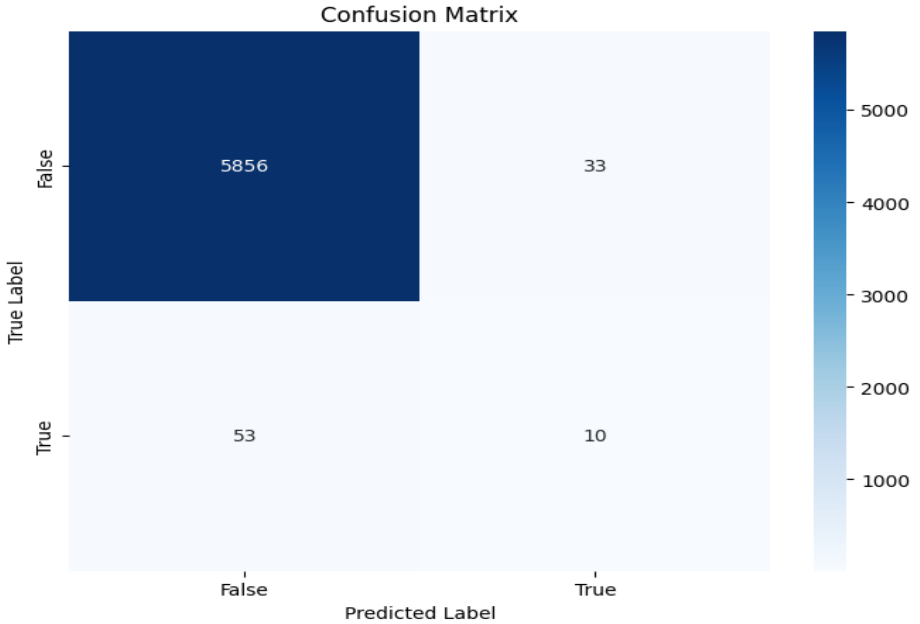
# SBERT vs. SRoBERTa Across Modalities

**Comparative metrics Graph** for Classification of Labels

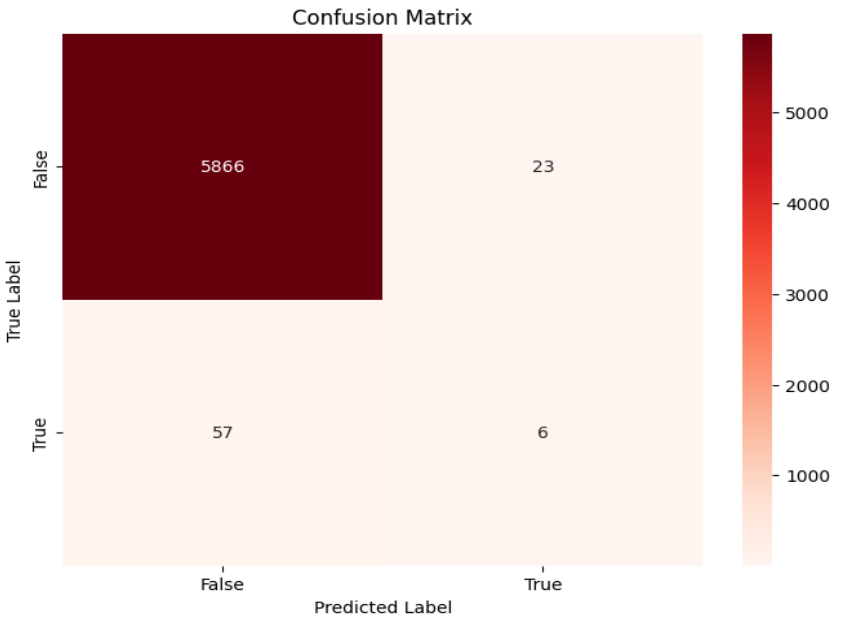


# Confusion matrix comparison

## Modality 1 - Balanced Split



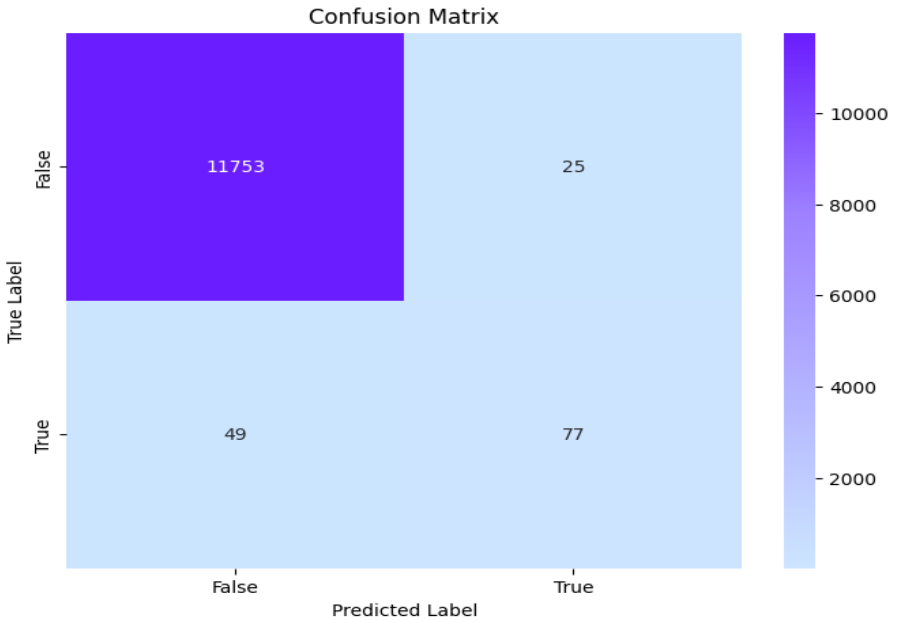

SBERT



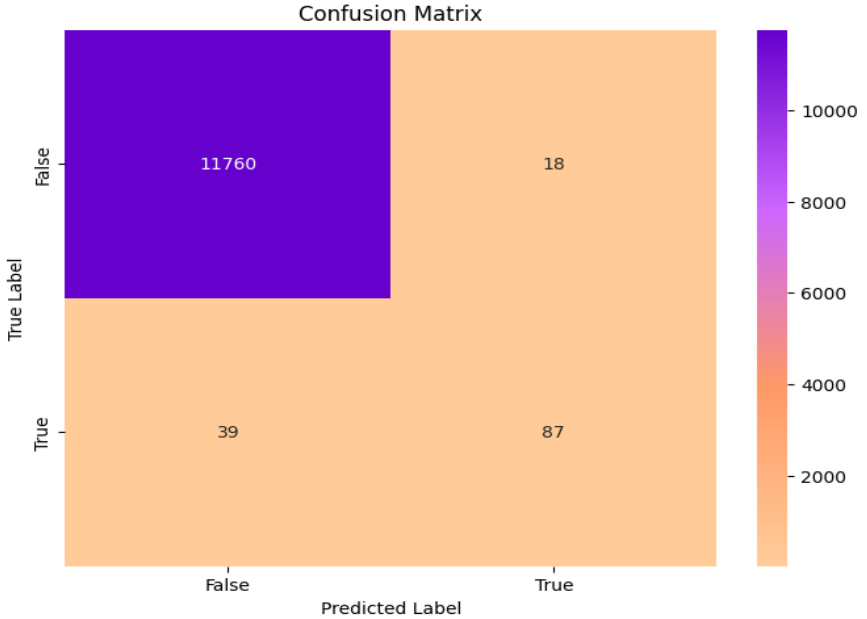

SRoBERTa

# Confusion matrix comparison

## Modality 2 - Dedicated Split




SBERT




SRoBERTa

# What’s next?

# Enhancing Model Performance and Capabilities

## Fine-tuning

- Optimize encoder hyperparameters to enhance the text semantics.

- Focus on retrieval tasks to extract the most relevant IDs and improve recall@k.

- Refine performance by TP > FN, for label predictions.

## Integrating Large Language Models

- Combine suitable LLM with the existing model pipeline for advance reasoning-based text generation.

- Leverage *contextual understanding* from the encoder to generate *explanations* for classified labels.

# THANK YOU!

**Grazie!**

**ধন্যবাদ**

**شكريه**