STATEC

Text classification with transformers

STATEC Datalab
Workshop

16/09/2025

Introduction

- **Official statistics rely on standardized classifications (e.g., NACE, ISCO or COICOP codes)**

- **Text-to-code problems : assign classification codes to textual descriptions**

  “I work as a software developer” ⇒ ISCO 21: Science and Engineering Professionals

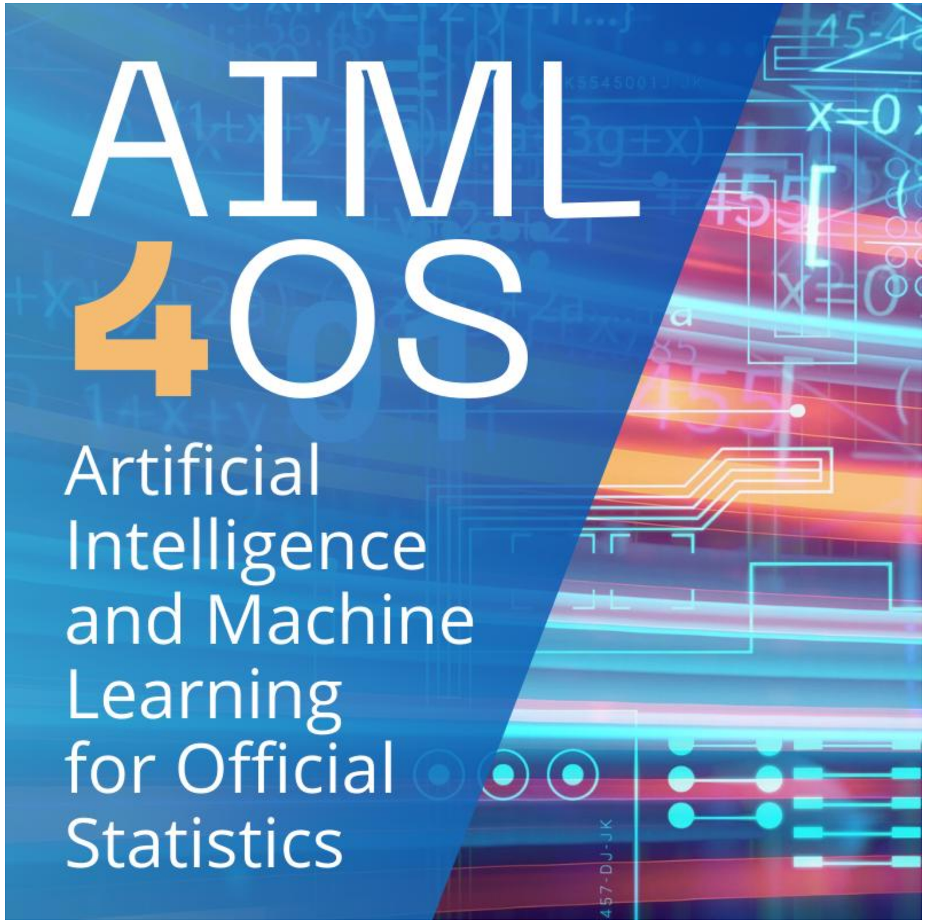
- **Manual coding can be time consuming**

- **Tutorials to demonstrate how transformer models can be applied to solve this problem more efficiently**

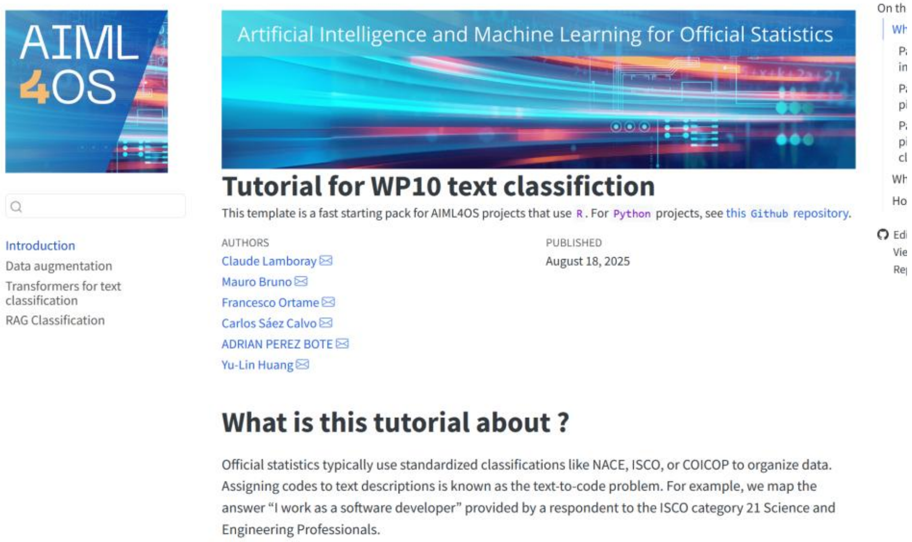
Introduction

- **European project AIML4OS**

  https://cros.ec.europa.eu/dashboard/aiml4os



Organization of the tutorial


AIML
4OS
Artificial Intelligence and Machine Learning for Official Statistics
Introduction
Data augmentation
Transformers for text classification
RAG Classification

Tutorial for WP10 text classifiction

This template is a fast starting pack for AIML4OS projects that use R. For Python projects, see this Github repository.

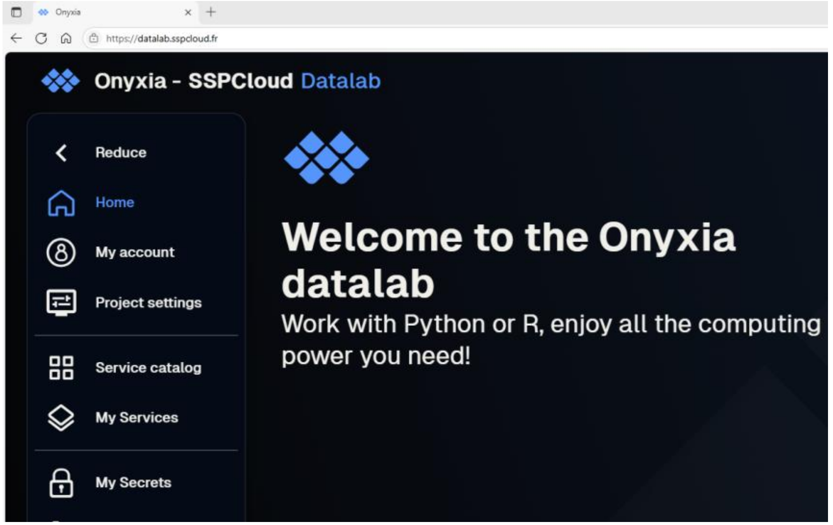
AUTHORS
Claude Lamboray
Mauro Bruno
Francesco Ortame
Carlos Sáez Calvo
ADRIAN PEREZ BOTE
Yu-Lin Huang

PUBLISHED
August 18, 2025

What is this tutorial about ?

Official statistics typically use standardized classifications like NACE, ISCO, or COICOP to organize data. Assigning codes to text descriptions is known as the text-to-code problem. For example, we map the answer “I work as a software developer” provided by a respondent to the ISCO category 21 Science and Engineering Professionals.

Launch Notebook on SSP Cloud



Welcome to the Onyxia datalab

Work with Python or R, enjoy all the computing power you need!

Notebooks on github website
https://aiml4os.github.io/WP10_tutorial_text_classification/

SSP Cloud as datalab platform
https://datalab.sspcloud.fr/

Need to create an account on SSP cloud
Contact innovation@insee.fr if email domain not covered

Agenda

1. **Synthetic data generation to improve the training data set**

Carlos Sáez Calvo, INE Spain

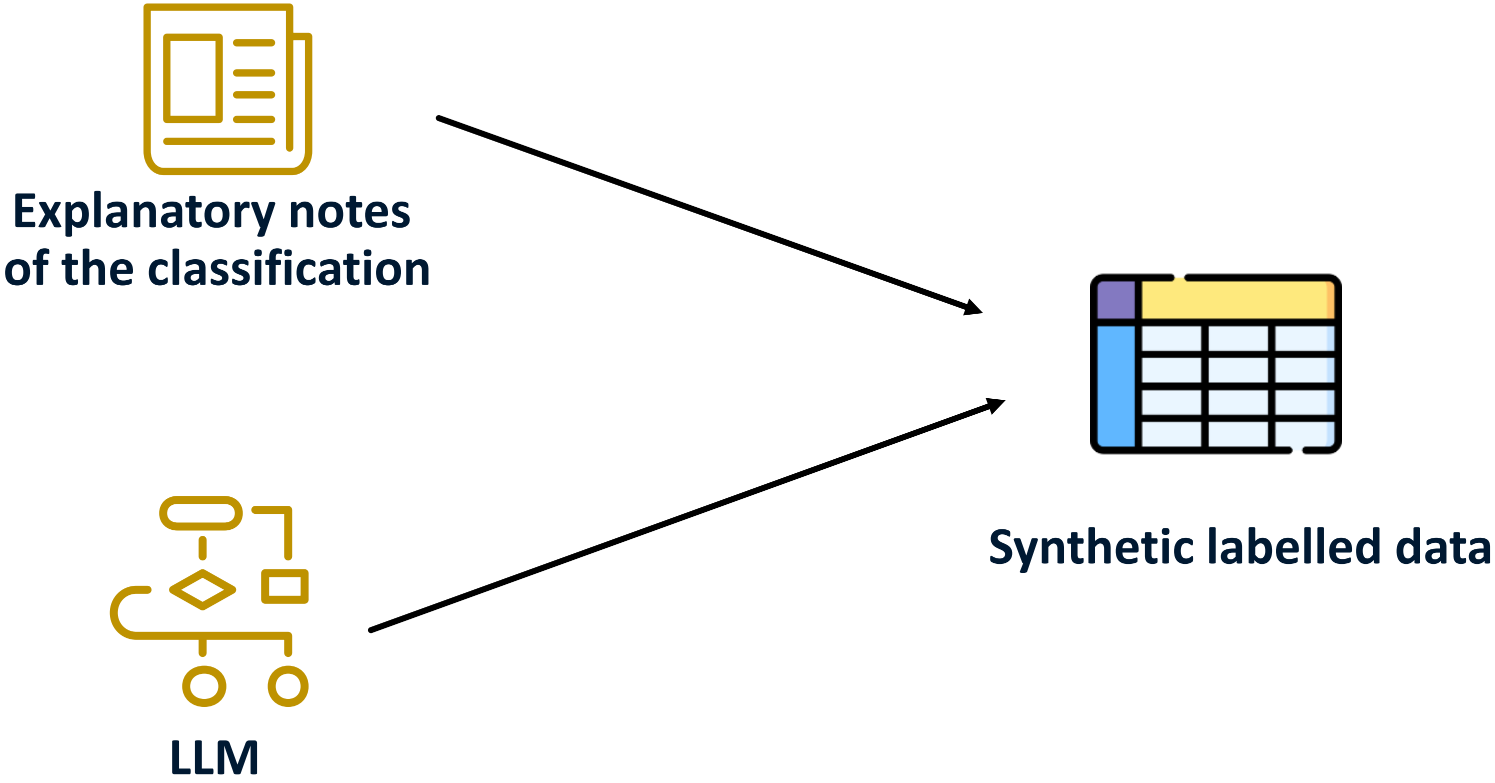
2. **Fine-tuning a transformer pipeline for text classification**

Yu-Lin Huang, STATEC and SnT (University of Luxembourg)

3. **Setting up a simple RAG pipeline for automatic text classification**

Francesco Ortame, ISTAT

1. Synthetic data generation to improve the training data set

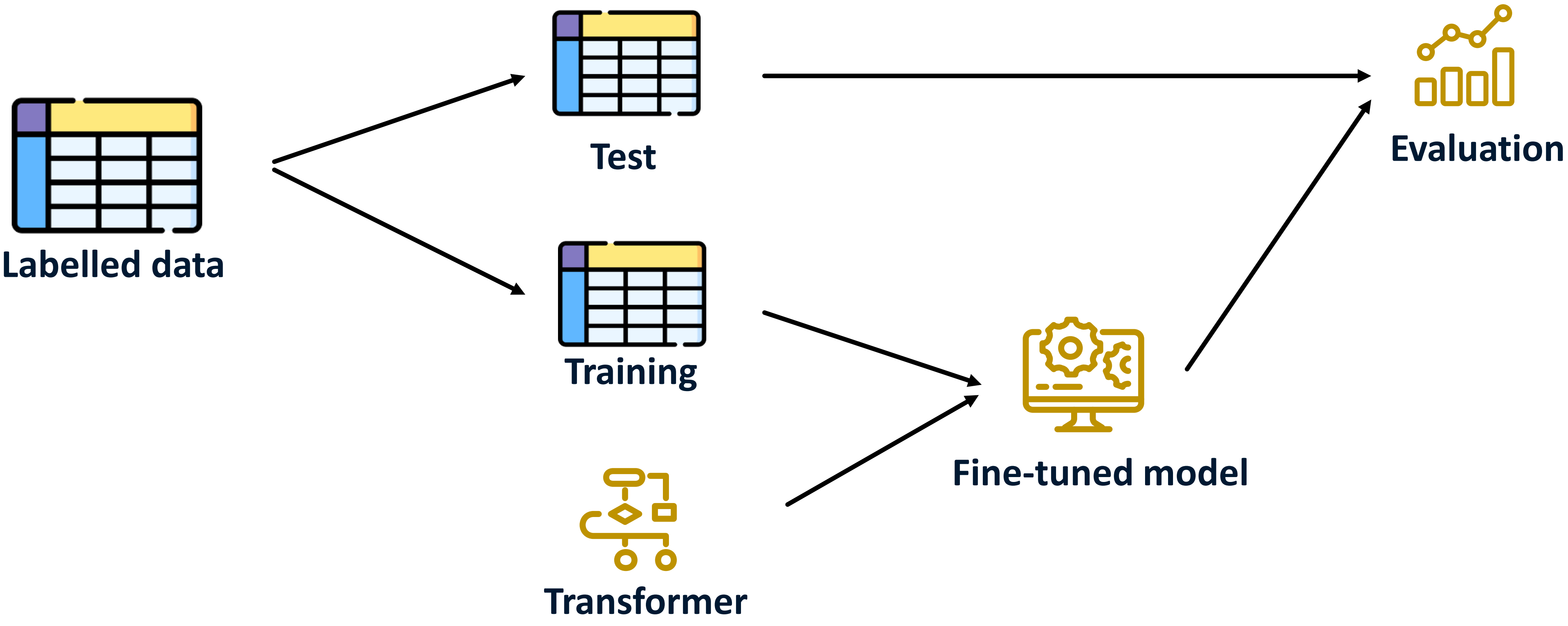


Explanatory notes of the classification

LLM

Synthetic labelled data

2. Fine-tuning a transformer pipeline for text classification


Labelled data

Test

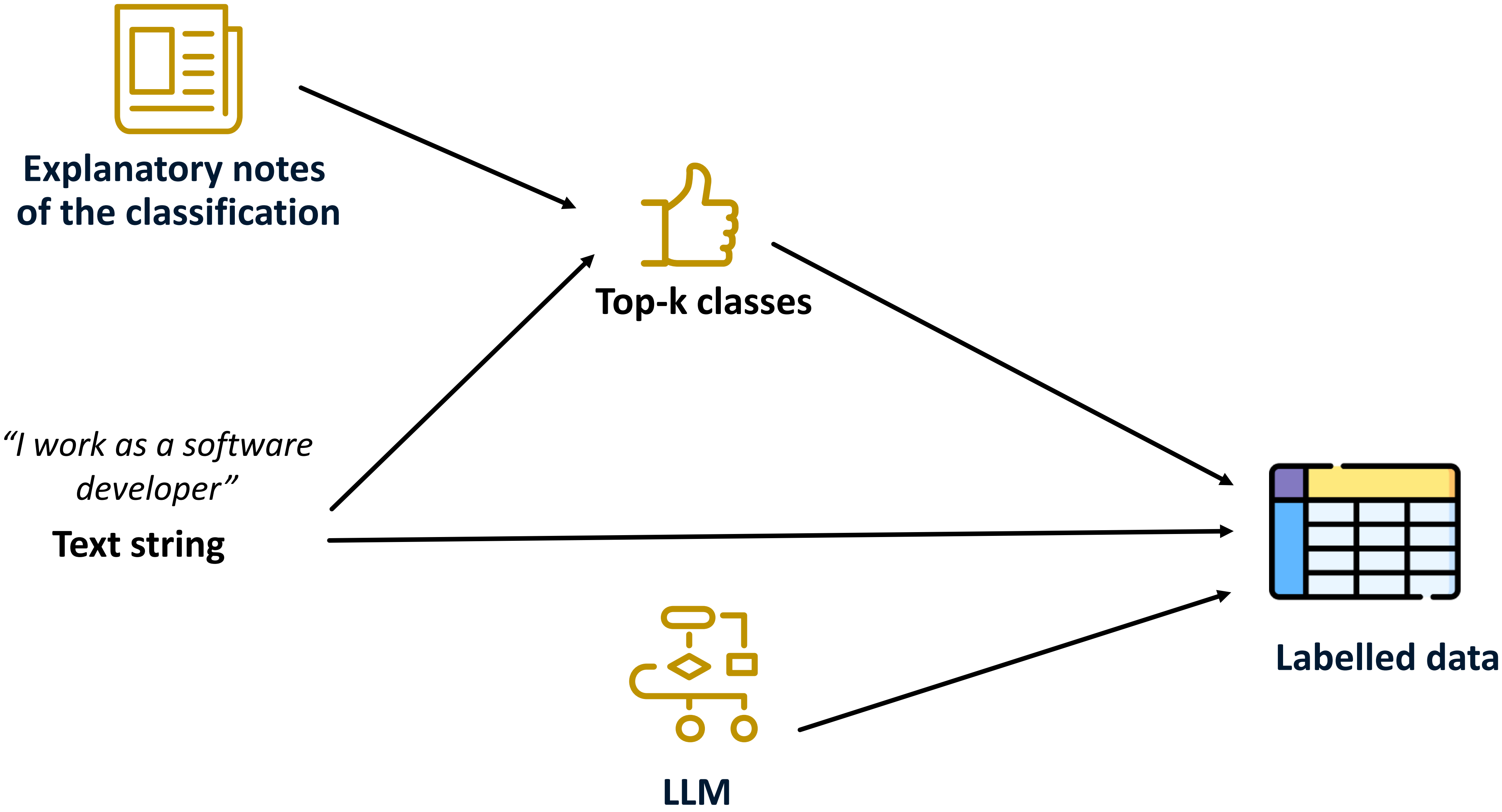
Training

Transformer

Fine-tuned model

Evaluation

3. Setting up a simple RAG pipeline for automatic text classification



Explanatory notes of the classification

Top-k classes

“I work as a software developer”

Text string

Labelled data

LLM