Retraining strategies for an economic activity classification model

European Conference on Quality in Official Statistics 2024

21 May 2024

Introduction

Machine learning in official statistics

Quality of ML systems

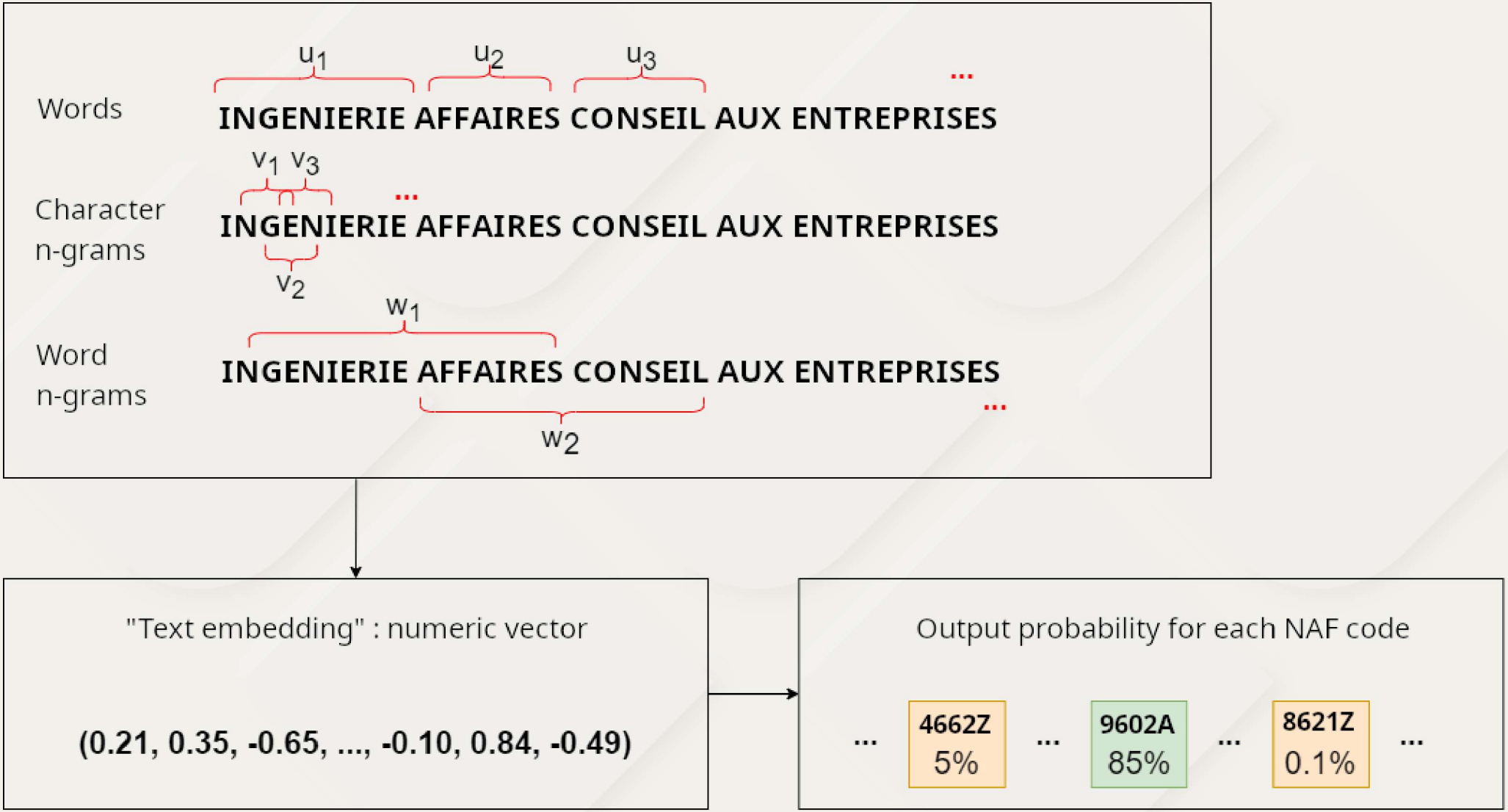
- A ML model is trained to solve a task **based on reference data**
- Real-life data **can deviate** from the reference data, which leads to performance issues
- **Retraining is necessary** to avoid these issues

Coding system

Description of the coding system

- Sirene is the French national **company register**
- When a company registers, an **activity code** is attributed
- A **model trained on historical Sirene data** is used when it is confident enough. Otherwise, the description is given a code manually

Modeling


Words
INGENIERIE AFFAIRES CONSEIL AUX ENTREPRISES
Character n-grams
INGENIERIE AFFAIRES CONSEIL AUX ENTREPRISES
Word n-grams
INGENIERIE AFFAIRES CONSEIL AUX ENTREPRISES
"Text embedding" : numeric vector
(0.21, 0.35, -0.65, ..., -0.10, 0.84, -0.49)
Output probability for each NAF code
4662Z 5%
9602A 85%
8621Z 0.1%

Performance

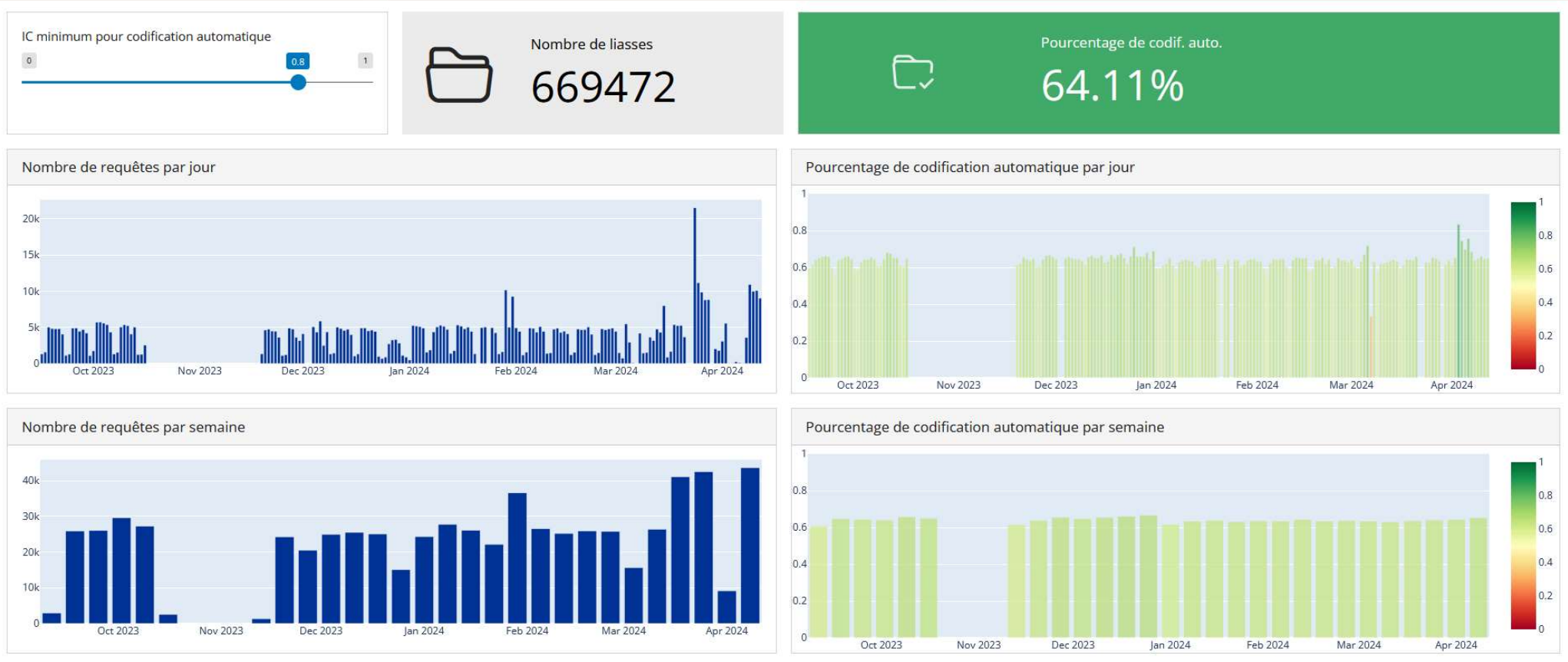
- Evaluation on historical data: **very high accuracy** of 89%
- With newer hand-coded data: **reduced accuracy** of 80% due to a distribution shift in the data
- **Company activities evolve over time**. New businesses appear, businesses traditionally associated with a certain activity may see this activity evolve, etc.

Monitoring

Design

- The model is served via a **REST API** (developed with FastAPI)
- A process fetches logs daily, parses them and **saves their content** on a persistent storage
- **An interactive dashboard** is built with Quarto to offer insight on data and how it is coded

Dashboard



Dashboard tab offering daily and weekly insight on the number of queries to the API and its automatic coding rate.

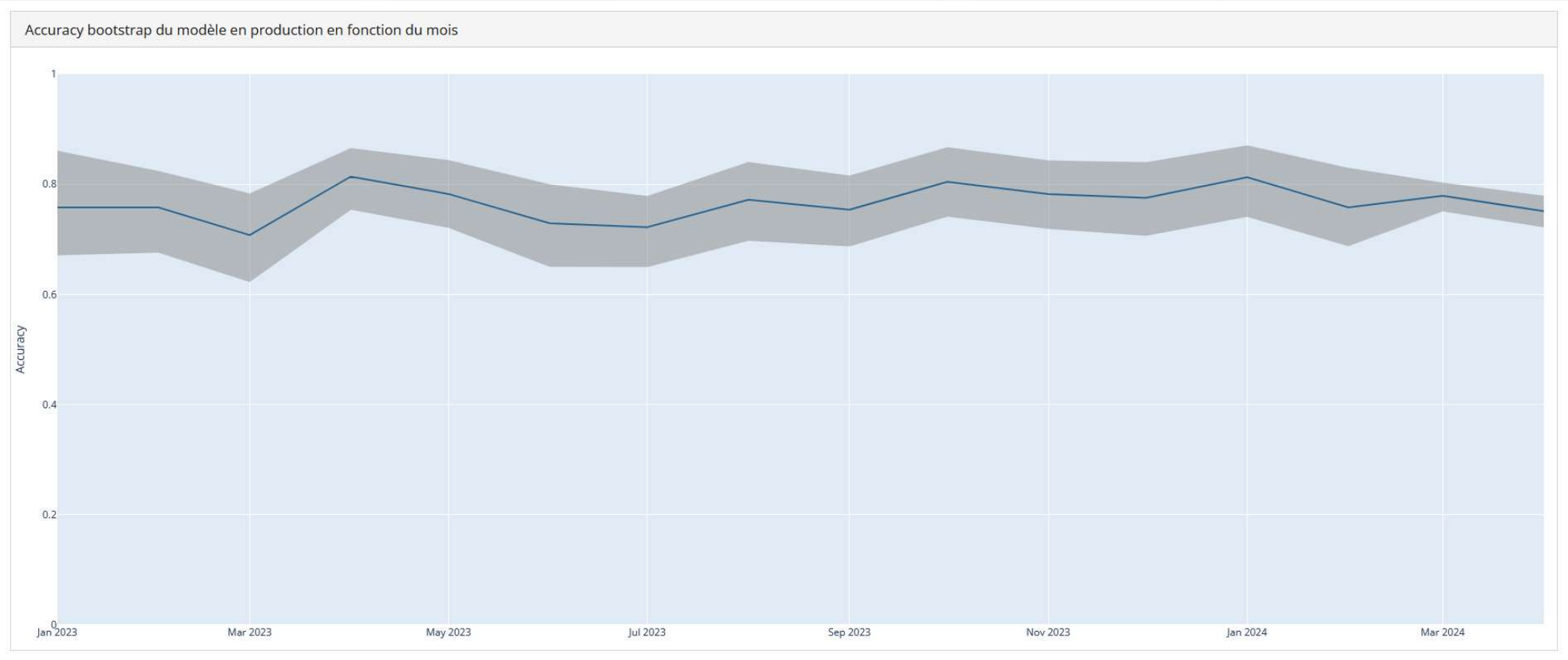
Dashboard

Dashboard tab displaying the two distributions of predicted classes at a specified level of the classification system for two specified time windows.

Continuous performance evaluation

- We continuously increment **an evaluation set** to monitor the performance of the ML system
- Batches are sampled **from recent Sirene data** and uploaded onto Label Studio
- For now data is **shared between annotators** and each description is coded once (could change in the future)
- The dashboard is enriched with additional tabs **leveraging evaluation data**

Continuous performance evaluation



Dashboard tab giving insight on the monthly accuracy of the evaluation set.

Retraining

Periodic retraining

- Company activities evolve over time. A first strategy is to **retrain the model periodically**
- How frequently ? There is a **tradeoff** as there should be a validation procedure to use a new model in production
- In our case, distribution shifts are not large. It is reasonable to **retrain twice a year**

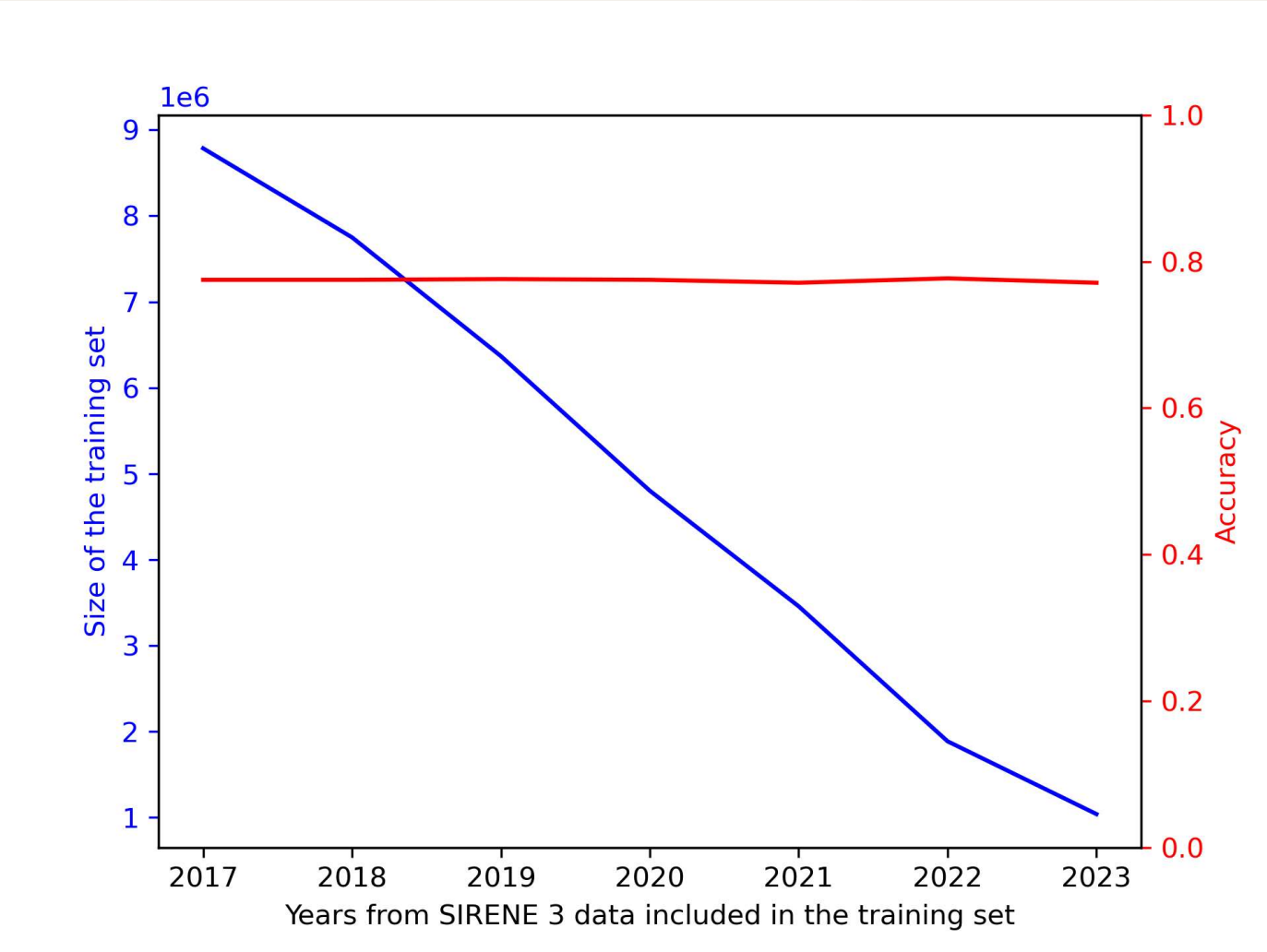
Additional retraining

- **Additional specific retraining procedures** can be triggered:
  - When the monitoring system **detects unusual shifts** in the data
  - When **repeated claims** are made by certain companies on their activity code
  - When **coding concepts change**

What training data ?

- When retraining a model from scratch: **how far should we go in the past** to build the training set ?
- **Empirical evaluation** is necessary:
  - Model capabilities **scale** with training data
  - Older data has **lower quality labels**

What training data ?



Accuracy and training set size as functions of the earliest year included in historical training data

Conclusion