Classifying companies in France using machine learning

UNECE Machine Learning for Official Statistics Workshop 2023

Thomas Faria, Tom Seimandi

5 June 2023

Context

Context

- **Several major changes**:
  - **Internal**: Revamping of the French company registry, **Sirene 4**.
  - **External**: Implementation of a **one-stop shop** to declare the creation of a business.
- **Observation**: **Sicore** is no longer a suitable tool ➡ 30% automated coding.
- **Consequence**: Ideal moment to propose a new methodology for automated NACE coding.

Data

- ≈ **10 million** observations from Sirene 3 covering the period 2014-2022.
- **Data labeled** both by Sicore and manually.
- An observation consists of:
  - A **textual description** of the activity
  - The **nature of the activity** – NAT (23 categories)
  - The **type of form** – TYP (15 categories)
  - The **type of event** – EVT (24 categories)
  - The **area (m^2)** – SUR (4 categories)

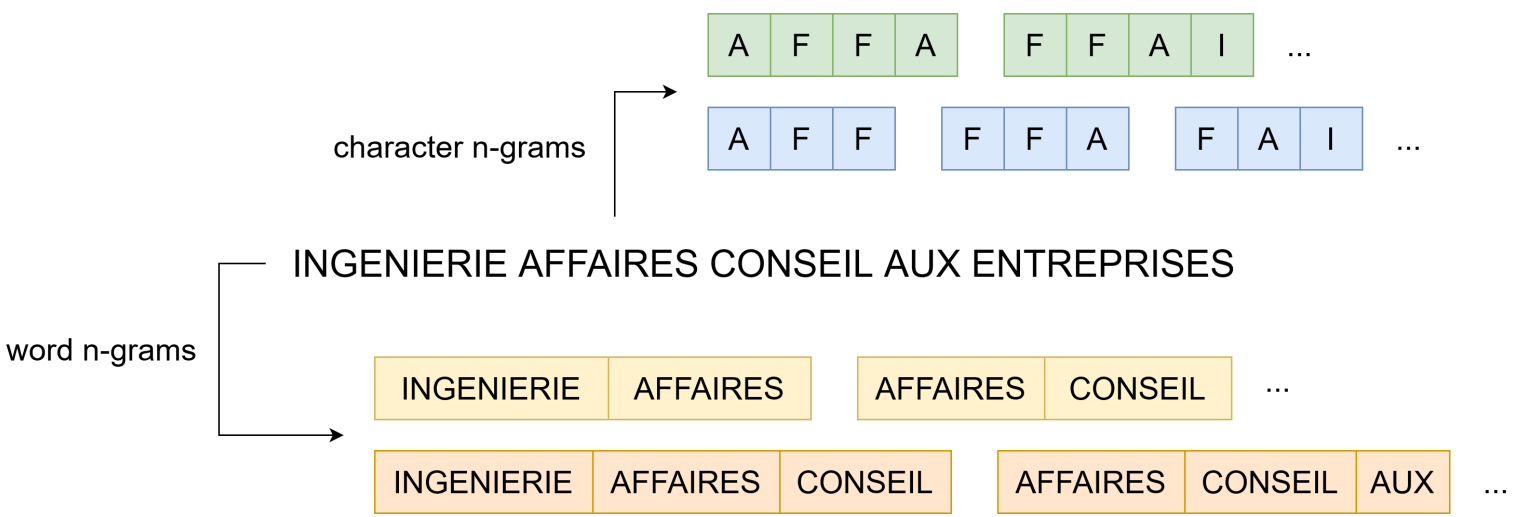
Hierarchical structure of NACE Rev. 2

Level	NACE	Title	Size
Section	H	Transportation and storage	21
Division	52	Warehousing and support activities for transportation	88
Group	522	Support activities for transportation	272
Class	5224	Cargo handling	615
Subclass	**5224A**	**Harbour handling**	**732**

Methodology

Feature extraction

- **Word embedding**: a method of **vectorisation**.
- **Pre-trained** embeddings available in open-source.
- We learn **our own word** embeddings.
- Additionally, embeddings for:
  - **word n-grams** and **character n-grams**.

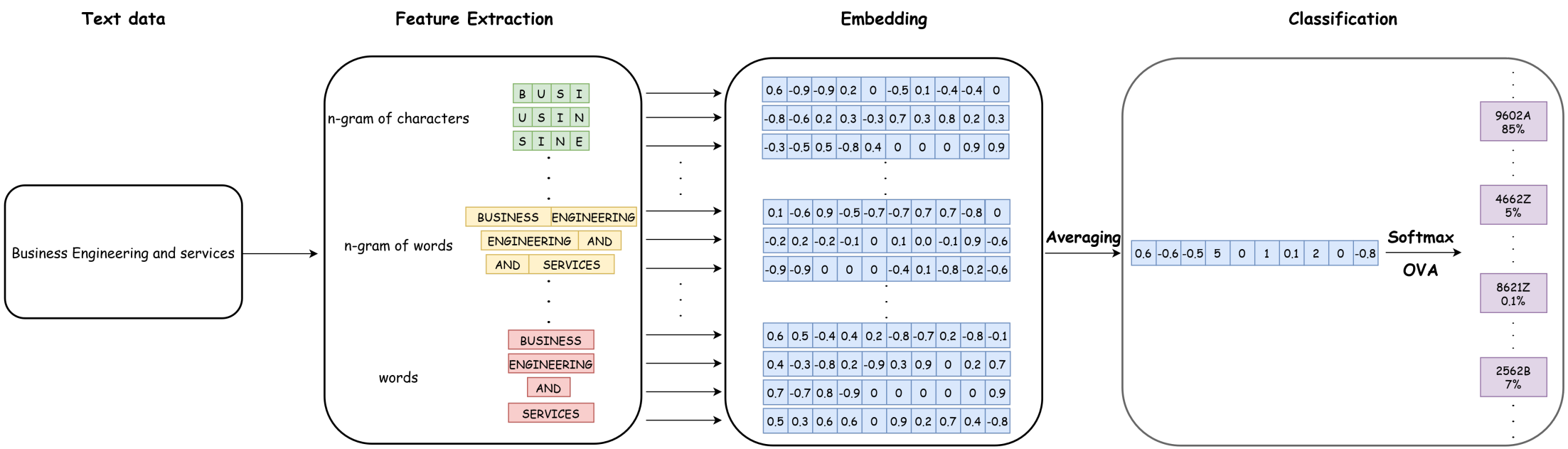


Linear classifer

- **2 classification methods**:
  - **Softmax**: a **single** multiclass classifier.
  - **One-vs-all**: **multiple** binary classifiers.
- **Optimisation**: **stochastic gradient descent** algorithm.
- **Loss function**: **cross-entropy**.

fastText model

- **fastText**: very **simple** and **fast** (C++) “bag of n-grams” model.



Handling categorical variables

- **Concatenation** of the text description with the names and values of the auxiliary variables:

Text	NAT	TYP	EVT	SUR
Cours de musique	NaN	X	01P	NaN

“Cours de musique NAT_NaN TYP_X EVT_01P SUR_NaN”

- **Imperfect method**: 3-grams “AT_” or “T_0” used.

Preprocessing

- **Preprocessing** essential for natural language processing.
- **Constraints**: **simple**, **light** and easily **reproducible** in Java  .

Transformation	Text description
Input	3 D: La Deratisation - La Desinsectisation - La Desinfection
Lower-case conversion	3 d: la deratisation - la desinsectisation - la desinfection
Punctuations removal	3 d la deratisation la desinsectisation la desinfection

Preprocessing

Transformation	Text description
Input	3 D: La Deratisation - La Desinsectisation - La Desinfection
…	…
Numbers removal	d la deratisation la desinsectisation la desinfection
One-letter word removal	la deratisation la desinsectisation la desinfection
Stopwords removal	deratisation desinsectisation desinfection

Preprocessing

Transformation	Text description
Input	3 D: La Deratisation - La Desinsectisation - La Desinfection
…	…
NaN removal	deratisation desinsectisation desinfection
Stemming	deratis desinsectis desinfect

Results

A good overall performance

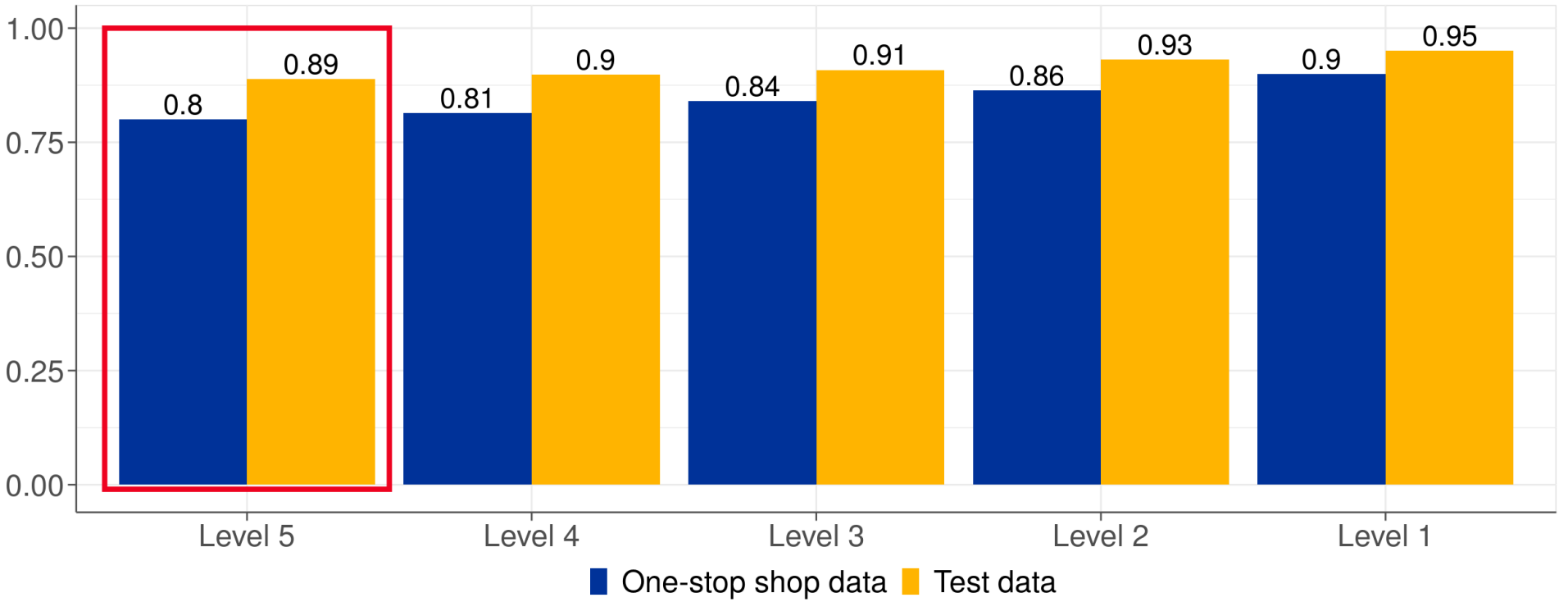


Figure 1: Accuracy for various level of the NACE nomenclature.

- Nearly **80%** of labels from the one-stop shop are correctly coded.
- Most prediction errors are **close** in the nomenclature.

Streamlining the manual coding process

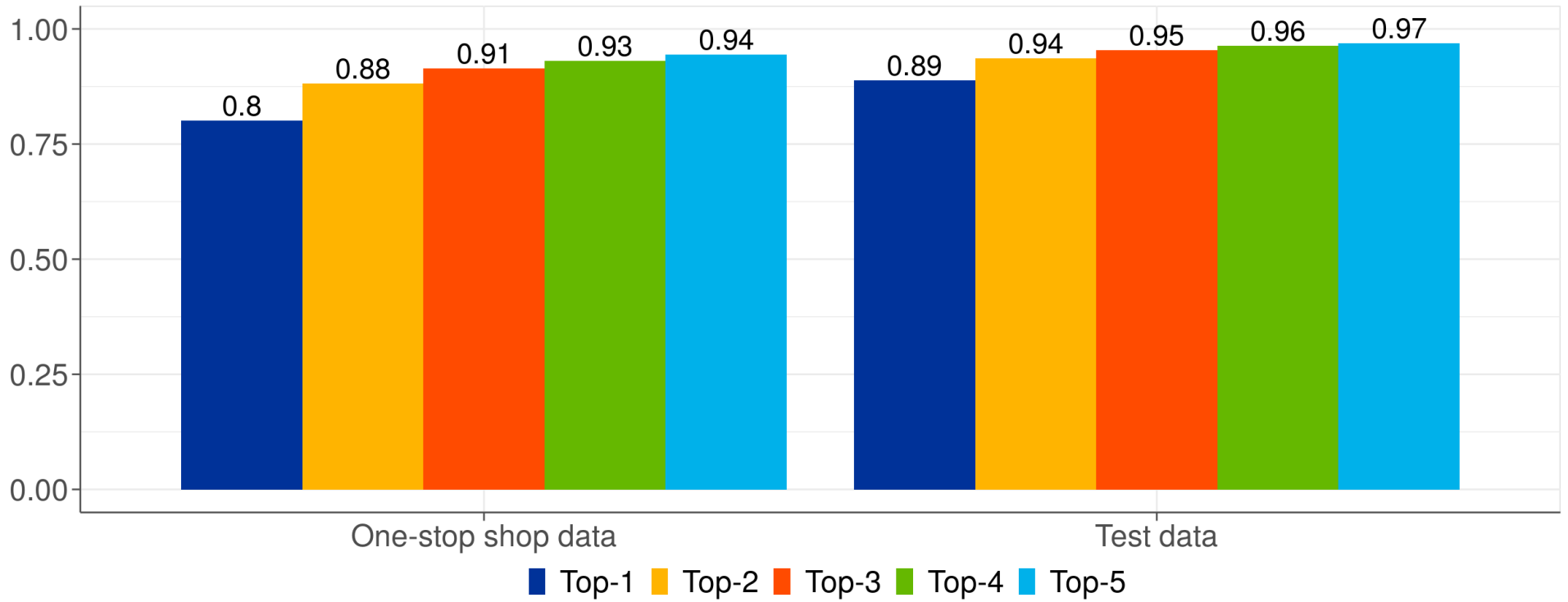


Figure 2: Top-k accuracy per sample.

Building a confidence index

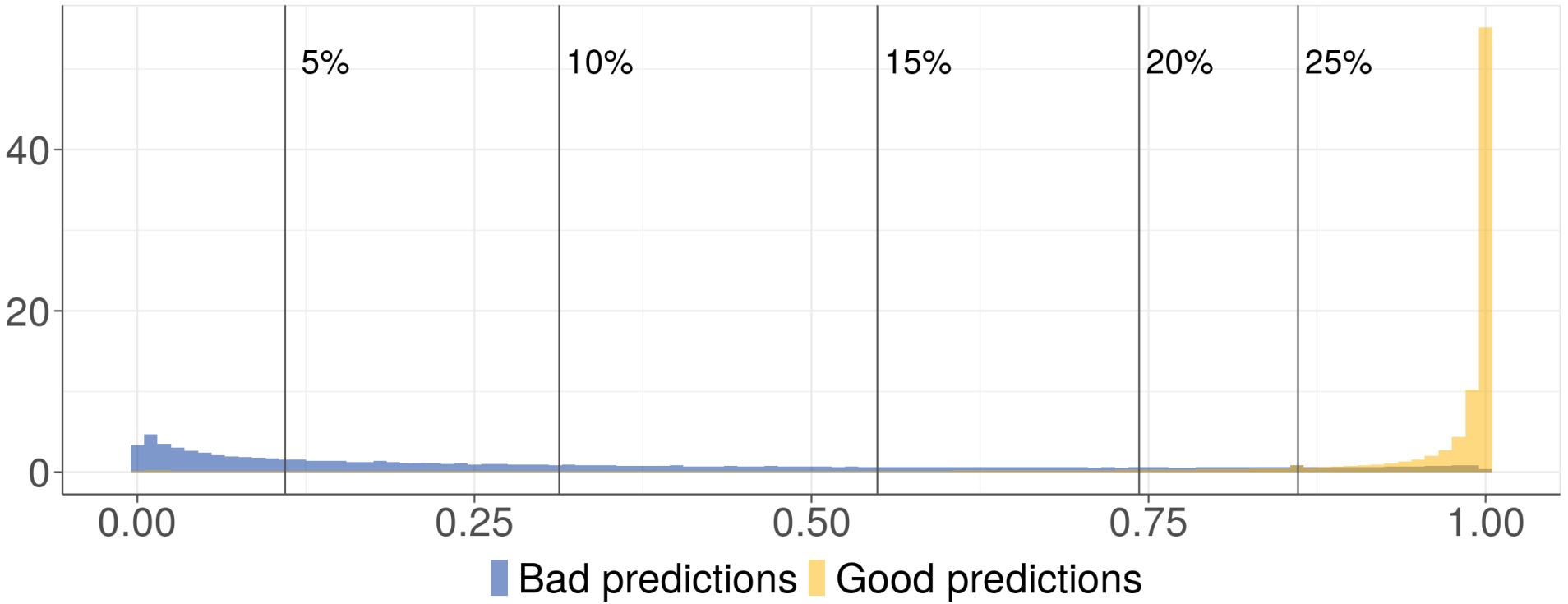


Figure 3: Distribution of the confidence index based on prediction results.

Efficiency of the manual coding process

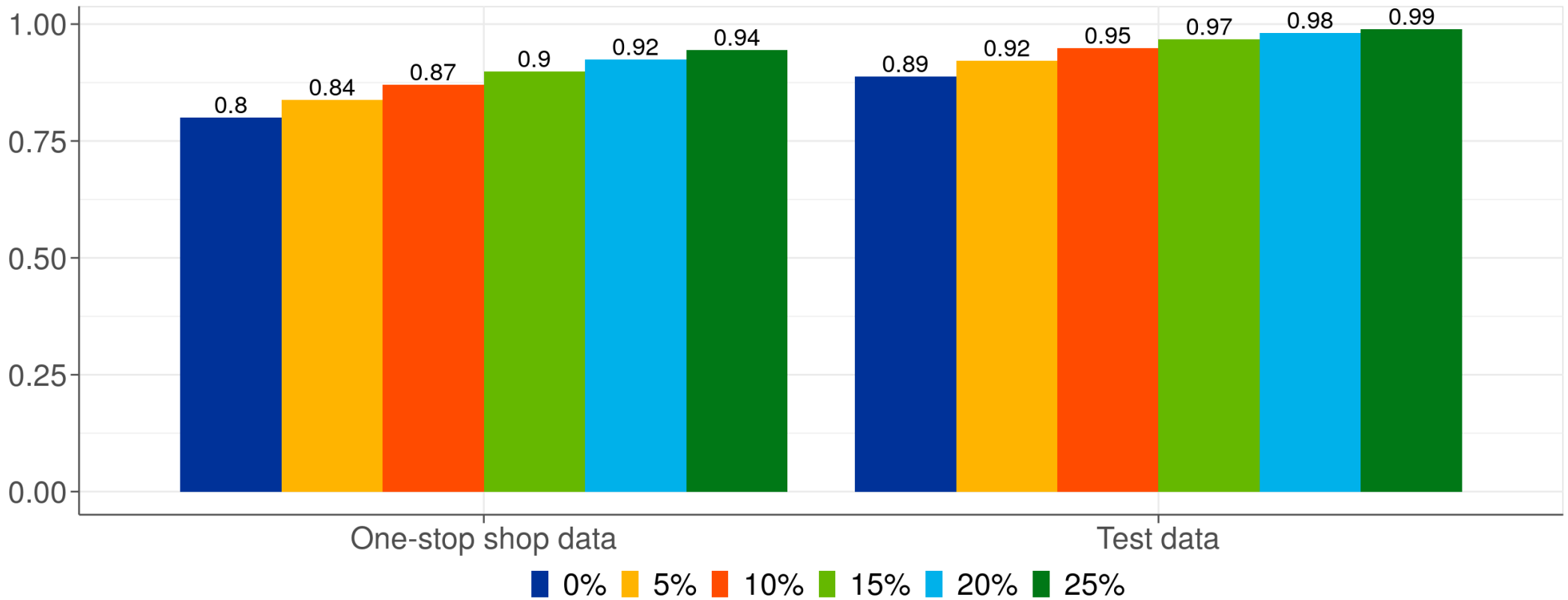


Figure 4: Accuracy for various shares of manual coding.

From experiment to production

Deployment of the model

- Models have been deployed in a **production environment** since November 2022.
- Emerging **new challenges** include:
  - **Organisational** issues.
  - Real-time **monitoring**.
  - Regular **re-training**.
- **MLOps** approach is required:
  - Join us on **Wednesday, June 7th** for a **Hands-on Lab**!