Classification of business activities by machine learning: The case of France

International Statistical Classifications Sprint, 2024
1 STATE OF PLAY: HOW WE ASSIGN AN ACTIVITY CODE
2 FASTTEXT: AN EFFICIENT MODEL FOR CLASSIFICATION
3 INSEE’S PLANS FOR NACE REVISION
STATE OF PLAY: HOW WE ASSIGN AN ACTIVITY CODE
HOW DO WE CLASSIFY THE ACTIVITY?

DIFFERENT PROCESSES

- Thanks to a literal description of the activity of the enterprise during an administrative procedure
  - The most important one: when an enterprise is set up, a form is filled in and a large amount of information is sent to Insee. This includes a literal description of the activity by the firm
  - New forms can be transmitted if the enterprise changes its activity
- Claims: if the enterprises disagree with the activity code given by Insee
- Surveys: structural statistical surveys (the so-called ESA, EAP) in which enterprises are asked about the breakdown of their activity
HOW DO WE CLASSIFY THE ACTIVITY?

FROM A LITERAL DESCRIPTION TO A CODE

– Up until November 2022, we used an automatic label coding system, called Sicore
  • Based on a training file of encoding examples
  • Drawbacks: if the label did not match an encryption example, no code suggestion was returned. It was then coded manually by a human being

– Since November 2022, we have implemented a new model based on machine learning: FastText
  • The training sample: 10 million observations coded by Sicore or manually
    ➔ Use of the literal description+auxiliary variables
    ➔ Need for preprocessing: lower case conversion, removal of punctuation, removal of numbers, removal of one-letter words, removal of stop words, stemming …
  • Very accurate even with literal descriptions that have never been coded before
  • A 100 % result even with a low accuracy rate
  • However we have decided to maintain a manual check if the accuracy is not good
HOW DO WE CLASSIFY THE ACTIVITY?

TWO POSSIBILITIES FOR HUMAN REVIEW: FASTTEXT OR MANUAL CHECK

Code APE incertain

Traiter via Fasttext  Traiter via la NAF

Informations sur l'activité:

Libellé d'activité:
travaux d'entretien, peinture, petits travaux de plomberie et électricité

Nature d'activité

Surface

Résultat Fasttext:

- 4334Z  Travaux de peinture et vitrerie
- 8121Z  Nettoyage ouvrant des bâtiments
- 9529Z  Réparation d'autres biens personnels et domestiques
- 3315Z  Réparation et maintenance navale
- 9609Z  Autres services personnels n.c.a.

Code APE sélectionné

Aucun code sélectionné actuellement
SOME RESULTS OF THE MACHINE LEARNING APPROACH

AUTOMATION RATE DURING 2023

Time Series of Automation and Human Review Rates: ML-assisted/Manual (NAF)
FASTTEXT: AN EFFICIENT MODEL FOR CLASSIFICATION
WHAT IS FASTTEXT?

FOR A RAPID AND LIGHTWEIGHT MODEL

- A versatile and efficient tool
  - for learning word representations
  - for sentence classification

- Developed by Facebook's AI Research (FAIR) lab, this open-source library is acclaimed for its swift processing of large datasets
  - Notably, it excels in rapidly classifying millions of data points
    - a task commonly challenging and costly for many machine learning models when handling complex tasks.
  - Particularly well-suited for company classification
    - seamlessly integrates into IT infrastructure without imposing excessive demands.
ONE-VS-ALL OR ONE-VS-REST CLASSIFIER

- **Approach:**
  - Train a binary classifier per class
  - Treat each class as positive, others as negative

- **Prediction:**
  - Obtain probabilities for each class
  - Choose class with highest probability

- **Advantages:**
  - Simple implementation
  - Probabilistic interpretation.
THE PRINCIPLE OF WORD EMBEDDING

- Dense abstract mathematical representations of individual words in a text, capturing context and surrounding word relationships.
FEATURE EXTRACTION: THE USE OF N-GRAMS

- Instead of applying ML model directly to the raw data for better results

Character n-grams:

```
AFFAF
F F A
FAI
```

Word n-grams:

```
INGENIERIE AFFAIRES CONSEIL AUX ENTREPRISES
```

```
INGENIERIE AFFAIRES CONSEIL
AFFAIRES CONSEIL AUX
```
A PRETTY LIGHT MODEL

- Simple neural network with only one layer
A GOOD OVERALL PERFORMANCE

Figure 1: Accuracy for various level of the NACE nomenclature.
SOME RESULTS OF THE MACHINE LEARNING MODEL

- Knowledge of probabilities for each class.
- The correct classification is among the top 5 most probable predictions in 94% of cases.

Figure 2: Top-\(k\) accuracy per sample.
MODELS HAVE BEEN DEPLOYED AND RE-TRAINED MULTIPLE TIMES IN A PRODUCTION ENVIRONMENT SINCE NOVEMBER 2022

Emerging new challenges include:

- Organisational issues
- Real-time monitoring
- Regular re-training
- Data annotation tool

MLOPS APPROACH IS REQUIRED:
INSEE’S PLANS FOR NACE REVISION
TWO MAIN ISSUES

- The flow: How to classify the activity of new enterprises according to the new NACE/NAF
  - Current classification: NAF 2008 (a more detailed version of the NACE Rev.2)
  - New classification: NAF 2025 (a more detailed version of the NACE Rev.1)
  - Consistent alignment required with both NACE and current NAF

- The stock: How to recode the activity of more than 10 million active legal units of our administrative business register (Sirene)
  - Not all textual declarations are retained in the historical database of firms for privacy reasons (deletion of too old data)
  - Not always direct correspondence between the old and new NAF
We plan to adapt our machine learning model

- Using data annotated manually by human experts
- Using some of the past data recoded to the new activity classification
- Retraining the model with updated output labels to ensure the method’s longevity

"Coding new enterprises: a two-stage process in both classifications"

- 2025: Major coding with NAF 2008, manually reviewed by human for complex cases. Minor coding with NAF 2025 with fully automated processing, followed by data quality operations for further improvement.
- 2026: Major coding with NAF 2025, manually reviewed by human for complex cases. Minor coding in NAF 2008 with fully automated processing, as the administrative business register officially adopts NAF 2025."
DIRECT CORRESPONDENCE BETWEEN BOTH CLASSIFICATION

- The correspondence table for the subclasses that are unambiguous
  - A one-to-one mapping between codes in the two classifications
  - Automatic recoding for over 40% of active legal units in the business register
  - 4.5 million of legal unit with an ambiguous code
  - (plus 3 million of renters of furnished accommodation, mainly households)
A NON-BIJECTIVE MAPPING BETWEEN CODES IN TWO DISTINCT CLASSIFICATIONS

1) Structural business survey (ESA/EAP) for the sampled enterprises (mainly the largest):
   - Activity codes for companies surveyed in 2023-2024 recoded based on questionnaire responses
   - We plan to add new product codes and some new questions

2) FastText applied to retained textual declarations (for around 70% of legal units with ambiguous activity codes)

3) Information from partner files for coding based on external data

4) Default code assignment for specific units registered as non-visible in the business register at the request of URSSAF or DGFIP.

5) Implementing default recoding methods in collaboration with sectoral managers in business statistics, including simplifications (automatically assigning the most frequent category for small businesses) or assessing the impact of potential changes in collective bargaining agreements to minimize disputes

6) Specialized survey to enhance the business register for complex and sensitive cases (has not yet commenced)
Nathan RANDRIAMANANA
Data Scientist in the Department of Registers, Infrastructures and Structural Statistics
Insee, Business Statistics Directorate
nathan.randriamanana@insee.fr