Automation of NOGA coding (NOGAuto)

**Keywords:** Machine Learning, Natural Language Processing, Artificial Intelligence, Prediction, Data Science, Data Innovation, Classification, NOGA and NACE codes.

# Introduction

In 2017, the Swiss Federal Statistical Office (FSO) released its data innovation strategy to ensure to keep up with time and technology. NOGAuto is one of the main five projects, which has been retained in line with FSO’s data innovation strategy with the goal to augment and/or complement the existing basic official statistical production at the FSO.

This project is realised to automate the coding of the economic activity of enterprises using “supervised machine learning” methods applied to already available data within the FSO (e.g. data from surveys, descriptions in the commercial register, explanatory notes for classifications) to support coding.

## Definition of classifications

Classifications are one of the basic elements for the production of statistics. The General Classification of Economic Activities (NOGA) is an essential tool for structuring, analysing and presenting statistical information. It enables the statistical units “enterprises” and “establishments” to be classified by their main economic activity and categorised into coherent groups. It can be used to depict the real situation as accurately and comprehensively as possible. NOGA 2008 takes into account both the framework conditions set by the statistical classification of economic activities in the European Community (NACE, rev. 2) and the needs of various interest groups in Switzerland.

The quality of the NOGA coding of the enterprises registered in the Swiss statistical business register (SBER) has a direct impact on the results of the structural, economic and synthetic statistics that concerns enterprises. Furthermore, this also highly depends on stable, monitored and high-quality NOGA coding of the SBER units. This is even more important in the context of register and administrative data, which are the starting point of countless official statistics, the base of sample frames and of statistical data analysis.

## The NOGAuto project

With a view to a continuous improvement of the quality in “units coding” in the SBER as well as to decrease the burden of businesses in their obligations to deliver information to the statistical offices, the FSO launched in 2018 a project called “NOGAuto” to automate the attribution of the economy activity code to businesses and therefore to an enterprise.

The coding procedures are currently fairly clerical. The employees who classify and codify the enterprises, analyse and interpret information based on the activities description of the businesses, such as data supplied by the businesses themselves, information from surveys, and descriptions from the company register and various administrative data. This is inevitably associated with a human and subjective interpretation of information available, which makes standardised coding difficult.

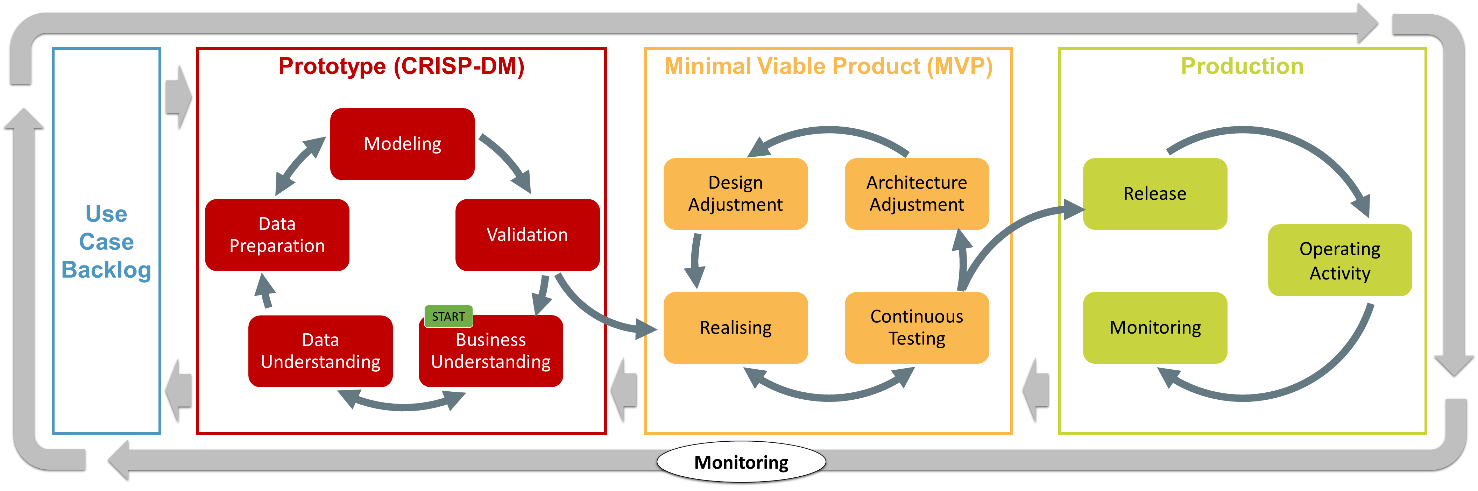
The coding procedures are currently fairly clerical. The employees who classify and codify the enterprises, analyse and interpret information based on the activities description of the businesses, such as data supplied by the businesses themselves, information from surveys, and descriptions from the company register and various administrative data. This is inevitably associated with a human and subjective interpretation of information available, which makes standardised coding difficult.

Therefore, the FSO is building a “machine learning system” to automate the manual coding procedure by using innovative technologies. This “learning machine” (i.e. artificial intelligence) undertakes the reading and interpretation steps from the coder and automatically classifies the company to a NOGA code. In addition, it proposes new keywords and concepts.

# Methods

In order for the project to be considered successful, it is important that it follows a predefined structure that will allow better control over the stages which have to be completed. Therefore, it has been decided to put in place 3 stages with different levels of success that the project has to achieve.

In the first stage (“Prototype”), the goal is to understand the business as well as the data, which will be used, and to begin the development and validation of the machine learning algorithms. In order to move from stage “Prototype” to stage “Minimal Viable Product (MVP)”, the system will have to achieve a success rate of at least 75% in terms of “generalisation error” with an acceptable time of response. In the second stage (MVP), the algorithms will be improved to a higher predictive accuracy of at least 85% in terms of “generalisation error” with an acceptable response time and an user-friendly interface, the project will move on to the final stage “Production” and be released and monitored.



An important role during all these stages is given to the different actors in this project in order to give their feedbacks and opportunities for improvement. The thematic experts as well as the users (i.e. the coders) have to give easily and continuously feedback depending on the stages of the project. In addition, documentation will be provided to ensure the best possible monitoring of the project.

We have carried out a pre-study on the feasibility of this project. The results obtained in this initial phase of the project are highly promising. This system (“NOGAuto”) performs the steps of reading and interpreting the descriptions of economic activities (“data preprocessing phase”) and predicts a category between the business and one of the 722 classification codes (“modelling phase”).

The first challenge of the “data preprocessing phase” is to detect the language. At first, the system has to determine the language of the description of the activity. German, French and Italian are the languages, which are the most spoken in Switzerland. Therefore, the first important step consists to identify the correct language and to allocate the according company to the right language dataset. In the second phase, “text mining” as well as a “text processing” are carried out, this means, for example, the punctuation, the numbers and the useless words (“stopwords”) are removed. Then the words are displayed in a matrix according to different “word embedding methods” in order to obtain structured data.

After the datasets have been evaluated and validated by the intern collaborators, some “supervised machine learning” methods will be used and performed on the data to get some prediction results about the NOGA classification (“modelling phase”). An important part of this task is to set the “parameters” and to tune the “hyperparameters” for each method. The latter help the model to find the most probable NOGA code for an observation. As already defined above, the system will deliver predictions on the classifications on the NOGA codes.

It is important to emphasis that the effort invested in realizing the NOGAuto project is not only limited to classification of the economic activities of businesses, but that a slight adaptation of the system would allow to wider use this technology to assign codes to, for example, professions, diseases, causes of death, products, within the FSO.

# Results

At the current time, there are no results (expect some promising preliminary results from prototyping). A first MVP is expected in the next months and an improved version will be presented during the NTTS 2021, together with highlighting the challenges to integrate it into production.

# Conclusion

To conclude we would like to mention the challenges we have to face:

* the detection of languages;
* “text mining /text processing”;
* identify the useful variables and related “feature engineering”;
* apply the “best” machine learning approach/method;
* ensure continuous feedback from the users;
* enable the monitoring of the NOGAuto once in production.

The objectives we have to reach include the following:

* creation of a tool that can predict the NOGA codes and link the enterprises registered in the SBER with a coding quality that is equal or superior to the manual coding currently carried out by the NOGA team.
* standardisation of coding and minimization of the “human interpretation” factor in the NOGA code allocation process.
* improvement of the quality of NOGA coding and consequently of the entire business statistics.
* in a second step, the horizon of the system could be enlarged by looking directly on the web for additional sources of information on the business to be coded.