Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Swiss Federal Statistical Office FSO
Division Interoperability and Registers

Automation of NOGA coding (NOGAuto)

1 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.

1.1 Definition of classification

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.

While in some other countries, the companies themselves can define the NOGA code, in Switzerland the FSO is in charge of this. 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. Although the NOGA code is primarily intended to serve as a stratification variable for statistical purposes, it is increasingly used by many administrative agencies for a variety of non-statistical purposes and has gained political prominence in the wake of the coronavirus pandemic.

NOGA coding must be stable, controlled and of high quality. 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 and of political and administrative decisions.

1.2 Project NOGAuto

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 entitled "NOGAuto"[1] to automate the attribution of the economy activity code to businesses and therefore to an enterprise.

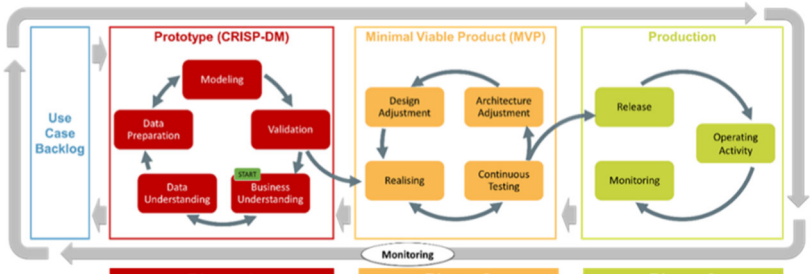
[1] https://www.experimental.bfs.admin.ch/expstat/en/home/innovation-data-science/nogauto.html

The coding procedures follow the same structure and are quite similar over the time. The employees, who classify and codify the enterprises, analyse and interpret information. These are based on activity descriptions of the businesses, such as data supplied by the businesses themselves, information from surveys, and descriptions from diverse registers 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 system) 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.

2 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 three stages with different levels of success that the project has to achieve.



		Phase 1	Phase 2	Phase 3
Prediction ML system-centered	Predictive accuracy (1 – generalisation error)	75%	85%	95%
	Documentation	✓	✓	✓
	Monitoring	✓	✓	✓
Domain user-centered	Interface	✗	✓	✓
	Performance	✓	✓	✓
	Daily use	✗	✓	✓
	Added value for the user	✓	✓	✓

In the first stage ("Prototype"), the goal is to understand the business as well as the data, which is 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 has 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 user 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 must be provided to ensure the best possible monitoring of the project.

The first challenge of the "Data Preparation" phase is to detect the language. At first, the system has to determine the language of the activity description. German, French and Italian are the languages,

which are the most spoken in Switzerland. Therefore, the first important step consists of identifying the correct language and allocating the according company to the right language dataset. For this reason, it was decided to create a "language detection algorithm" using a combination of diverse language R packages followed by "root cause analysis" methods.

In the second step of the "Data Preparation" 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" to obtain structured data and to extract the most information as possible.

After the datasets have been evaluated and validated by the intern collaborators, some "supervised machine learning" methods are used and performed on the data to get some prediction results about the NOGA classification (this corresponds to the "Modeling" 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 delivers predictions on the classifications on the NOGA codes.

It is important to emphasis that the effort invested in realising 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. This project could also be a potential supporting tool in the double codification of the next NOGA revision for the complex cases namely companies with 1:n relations.

3 RESULTS

Once the models were trained with the "Gradient Boosting Machine" (GBM) method in the prototyping phase, the results were verified and validated by the coding experts. After having validated them, the project moved on to the second phase, MVP. There the different algorithms were improved and an "User Interface" (UI) has been set up so that the coding experts can use it as a new working tool. Thanks to the close collaboration with the coding team, the prediction system was then fine-tuned, so that it allows simple activity descriptions to be coded and enables the coders to concentrate on the difficult ones with guidance.

4 CONCLUSION

The objectives we must 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 minimisation 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.