REGIONE DEL VENETO

InnovatiVE
SMART SPECIALISATION STRATEGY
VENETO

S3 Monitoring-System Veneto region 2021-2027

Classification document

innoveneto.org

Index

Introduction

Context

The Smart Specialisation Strategy (S3) is the tool that, since 2014, the EU Regions and member States must adopt in order to identify objectives, priorities, actions capable of maximising the effects of investments in research and innovation, by focusing resources on the specialisation areas of each territory.

To answer the needs of accounting, transparency and public information regarding the S3 and its application, Veneto Innovazione, in collaboration with Siris Academic, has published an online platform that provides institutions, companies and citizens with a wealth of 'open data' on research and innovation policies and the projects they have funded and implemented. The platform, divided into pages with different purposes, allows consultation, by search filters, of various indicators such as the number of projects funded, millions of euro in grants received from the actors of the regional ecosystem.

The data used in the platform regard projects developed in the Veneto region and/or by actors from the region. These projects, when collected from their sources, can either be already classified according to the Veneto region's S3 for the period 2021-2027 or not. In the latter case, there is the need to provide a trustworthy method to classify the projects in the Veneto's S3 framework. In this context, it has been decided to develop and implement an automatic classification system.

Deep learning approach

A classification system, based on deep machine learning techniques, uses algorithms which rely on neural networks. These neural networks, in case of a text-based classification, are applied to textual data (human language).

Thanks to the scientific developments of recent years in the field of natural language processing (NLP) and the open availability of models trained on huge amounts of data, this technique allows a better understanding of the description of projects, going beyond the words used and better capturing the context.

The results that can be obtained have high accuracy, but the perimeter must be defined concisely with a set of relevant concepts or with a seed corpus set representative of the perimeter set. By taking advantage of training models on huge text collections, this technique is able to find more complex and non-explicit semantic relations in the text.

Data classification workflows

Phase 1: Perimeter conceptualisation and definition

Definition of categories

The project, for which automatic classifier models have been developed, consists in the design and development of a regional S3 monitoring system for Veneto Region. In this context, the first step of phase one has been the collection and analysis of the available material and documents of Veneto's Smart Specialisation Strategy to identify the main categories of the S3. The documents collected, studied and analysed are:

- Modello di Monitoraggio e Valutazione della Strategia di Specializzazione Intelligente (S3) della Regione del Veneto 2021 - 2027 (Dgr_1684_22);
- Strategia di specializzazione Intelligente (S3) della Regione del Veneto 2021 2027 (Dgr_474_22_AllegatoA).

According to the analysis of these documents have been identified the following type of categories:

- Ambiti → 6 main areas of specialisation of the S3 structure, divided in 52 trajectories;
- Missioni Strategiche → 2 areas representative of the overall objectives of the S3, transversal to the 6 ambiti.

Specification of categories

The following step of phase one of the classification workflow has been the specification of the characteristics and perimeters of the Ambiti, according to the analysed documents' definitions of them and of their sub-categories (for the Ambiti).

The Ambiti and Missioni strategiche have been defined as follows:

- Smart Agrifood (Ambito) → According to the definition of its 10 trajectories;
- Smart Manufacturing (Ambito) → According to the definition of its 11 trajectories;
- Smart Health (Ambito) → According to the definition of its trajectories;
- Cultura e Creatività (Ambito) → According to the definition of its 10 trajectories;
- Smart Living & Energy (Ambito) → According to the definition of its 12 trajectories;
- Destinazione Intelligente (Ambito) → According to the definition of its 5 trajectories;
- Bioeconomy (Missione strategica) → According to its own definition;
- Space Economy (Missione strategica) → According to its own definition;

Phase 2: Classification methodology development

To effectively implement our classification process, it is crucial to have a well-structured training dataset containing rich examples for each category, especially examples on the decision boundaries. This training dataset allows the model/algorithm to extract patterns and learn the nuances of the text, facilitating the accurate prediction of category labels. The training set provides the model with a representative sample of labelled projects and publications to refine its understanding and improve its performance.

Additionally, an external test set is essential for evaluating the performance of our automatic labelling methods. This test set is composed by regional projects and labels provided by project beneficiaries, separate from the training data, is used to assess how well the model generalises to new and unseen projects. By comparing the predicted labels against the actual labels in the test set, we can measure the accuracy and reliability of our automatic classification. This evaluation ensures that the model not only performs well on the training data but also maintains high accuracy in real-world applications.

Training data

Data collection and "training" dataset creation

This step began with the collection of raw data to be later used for the creation of the dataset. Once we gathered the data listed, we cleaned them and selected only selected those that were relevant for us, according to the matching or not of these characteristics:

- We translated to English those in Italian, because language models for R&I text are only available in English
- We combine title and abstract

Thanks to this selection, we obtained:

- 5,000 European projects
- 5,000 Publications
- 2,800 Interreg projects

These documents have been stored and utilised for the creation of the dataset that we later used to train our automatic classifier models.

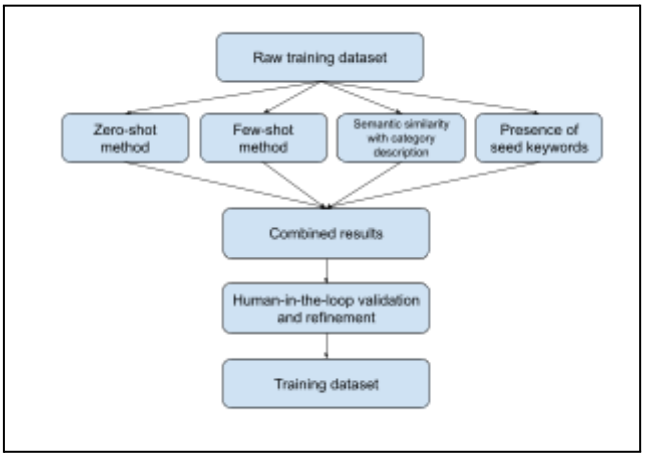
Labelling samples automatically

The labelling process for our dataset employs an ensemble of automated methods to enhance accuracy and robustness in categorising data into 8 priority categories for the region of Veneto (6 Ambiti and 2 Missioni strategiche). Our ensemble of methods use:

1. First, the **zero-shot approach** leverages pre-trained language models to predict category labels without any task-specific training only with category name, allowing us to handle unseen data effectively. We use the SCIroShot method[1], which is adapted to categorise scientific document according to R&I categories.
2. Next, the **few-shot method**[2] is employed, where a small number of labelled examples extracted from journals with similar names to categories of interest guide the model in learning the category distinctions, thus improving its performance with minimal labelled data.
3. Additionally, we apply **semantic similarity** with a scientific language model[3] with definition of trajectories in Ambiti and Missioni strategiche, where we measure the textual similarity between the projects and publications and category descriptions, ensuring that the labels are semantically aligned with the category meanings.
4. Lastly, the presence of **seed keywords**[4] method involves identifying key terms strongly associated with each category. Their presence in the data points significantly influences the labelling decision.

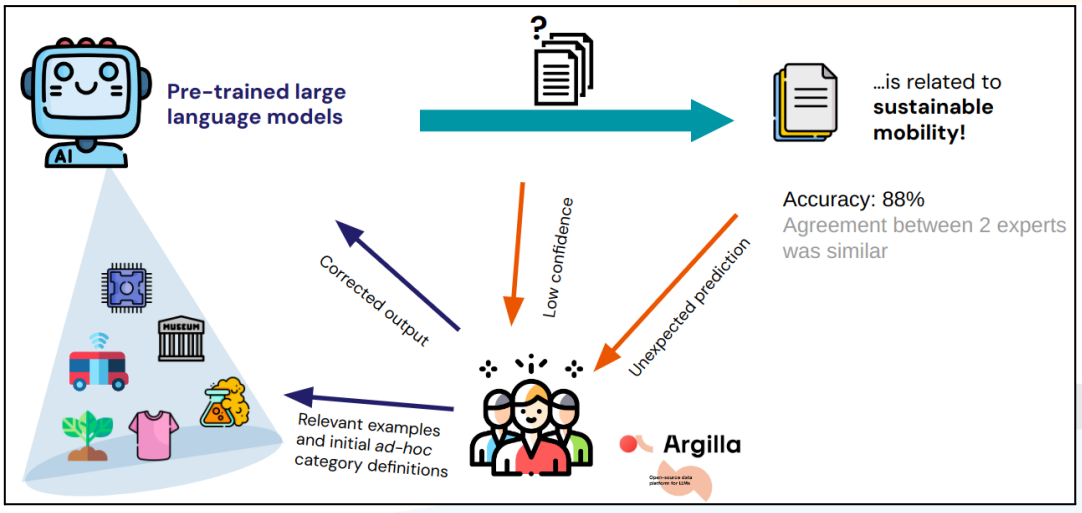
By integrating these methods, we achieve a comprehensive labelling strategy that combines the strengths of each approach, leading to more accurate and reliable categorisation into our five predefined categories.

[1] https://github.com/bsc-langtech/sciroshot
[2] We use the SetFit framework.
[3] https://github.com/allenai/scidocs
[4] Keywords described in this spreadsheet.



Human in the loop

To enhance the accuracy of our labelling process, we incorporate human-in-the-loop methods to clean and validate samples. Specifically, we manually review and verify those data points with lower probability scores, ensuring the correctness of the labels according to the specific definition of the trajectories from the perspective of the region of Veneto. This iterative process involves several rounds of cross-validation, utilising the Argilla tool to systematically refine the dataset. By combining automated techniques with human expertise, we achieve a higher level of precision and reliability in our dataset labelling.



Automatic classifier

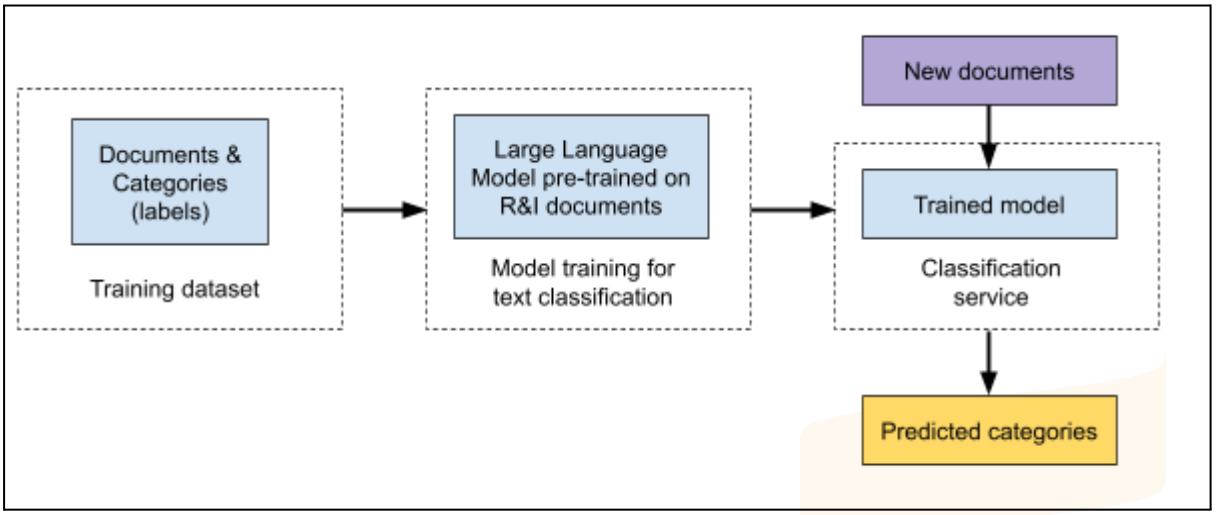
Training process

We train a text classification model on a LLM pre-trained on R&I document called *Specter*[5]. Our training data is described in the previous section. We train a classifier

[5] https://github.com/allenai/scidocs

that given a title+description of project/publication can predict the defined 8 categories. Training data contains projects and publications from different territories, with the aim of having a better generalisation of the model across different regions and contexts, beyond what is currently researched in Veneto, but according to the scope of categories defined by the region. This ensures its applicability to different regions for benchmarking purposes.

We train a text classification model with fine-tuning approach[6], adjusting the model's parameters based on the specific characteristics and nuances present project-label pairs in the training set.



Following the fine-tuning process, we conduct testing on regional projects to evaluate the effectiveness of the refined training dataset. This testing phase helps us assess how well the model performs in accurately categorising and analysing projects.

Phase 3: Evaluation of the results

Evaluation data

To evaluate performance of our automatic classifier, we used regional projects labelled by the beneficiaries themselves according to the areas. However, we identify inconsistencies in the category assignation, probably due to the crowdsourced nature of the annotations. This inconsistency leads to discrepancies in the understanding of each category's scope, and some projects may not be properly labelled according to their goals and missions.

Despite these challenges, this dataset provides valuable insights into the diversity and quality of the automatic categorisation efforts. Comparing automatic categorisation with labels provided by beneficiaries allows us to identify areas for improvement and refine the classification method. Additionally, it highlights the importance of providing clear guidelines and instructions to the beneficiaries when labelling projects to

[6] Method described in https://aclanthology.org/N19-1423/.

minimise discrepancies and ensure alignment with the intended categorisation framework.

Evaluating the performance

To evaluate the performance of the developed classifiers, we compared the regional projects matching similarity between the labels provided by the project coordinators and the ones provided by our automatic method. From the matching of these two methods we obtained the following results for the 8 categories:

	Smart Agrifood	Smart Manufacturing	Smart Living & Energy	Cultura e Creatività	Smart Health	Destinazione Intelligente	Bio economy	Space economy
Percentage of matching (both positive and negative)	98.16%	93.48%	90.78%	92.20%	96.88%	92.34%	97.16%	92.77%
Percentage of matching positive	88.68%	89.63%	85.79%	79.37%	76.27%	67.95%	26.32%	52.05%
Percentage of matching negative	98.93%	95.47%	92.62%	94.99%	98.76%	95.37%	99.13%	97.47%

The classifiers' performance data demonstrate that a high level of precision has been achieved.

The worst performances, which are still at an acceptable level, have been registered in the classifiers of the two Missioni Strategiche, as expected. These lower performances are due to the scarce definitions of the Missioni, especially if compared to very detailed definitions of the Traiettorie, which compose the Ambiti.

The actual trustworthiness of the automatic classifiers may be even higher than the statistical results indicate. This is because the labels of the regional projects, which we used for comparison, were assigned by the project coordinators who are not experts in the S3 domains. Below are examples of projects believed to be mistakenly classified by the project coordinators.

Examples of automatic classification

Title	Description	As classified by project coordinators	Result of the automatic classification method
Atlantic Terme Hotel: sustainability inclusiveness and digital innovation to improve performance	Atlantic Terme Natural Spa & Hotel is a superior 3 -star hotel located near Abano Terme. The structure is equipped with refined and high quality rooms and can count on the two thermal swimming pools in natural stone, a powerful leverage of attractiveness for Italian and foreign tourists. With this project, the proponent intends to invest in the perspective of: environmental transition (installation of photovoltaic system and replacement of obsolete and not very performing windows); of technological and digital innovation (development of a new site with online booking system equipped with multilingual translations) and greater accessibility (replacement of the flooring, track to mobilize people with reduced mobility and mobile lifter for swimming pool). The aforementioned investments will introduce tools, systems and plants that are not present in the structure and which will therefore be functional to the innovation of the structure and the services offered to customers, who will benefit from a smart, inclusive and attentive environment at 'environment. non_pertinent	Smart Living & Energy,	Smart Living & Energy, Destinazione Intelligente
Green and	The project provides for the redevelopment of the refuge in a green key with	Destinazione	Smart Living & Energy,

accessible refuge	particular attention to the subjects of disabilities. In fact, the installation of a photovoltaic system is expected, the creation of a bioclimatic pergola (partly dedicated to disabled subjects) and the purchase of new kitchen equipment that guarantee maximum energy saving non_pertinent	Intelligente	
ALTERNATIVE SKILLS 4 H2	Accustinies Srl is an innovative "eco-industrial" vocation startup, specialized in the sectors of electrochemistry, basic chemistry, metalworking, mechatronics, for the controlled production of hydrogen, understood, in the wide sense. The startup has conceived, developed and patented the Hymoov device for the production of hydrogen and oxygen on-demand (without storage), to be installed as a retrofit on any internal combustion engines (MCI) to improve their performance, to reduce CO2 emissions And the pollutants, to optimize the combustion of the engine and to reduce consumption (4 patents deposited of 2 internationals). The device is part of a transitional phase in which the current engines are transformed into less polluting and less "energetic" engines thanks to the production of hydrogen in injection (HFI) on-demand through the electrolysis process while the combustion engine is "in march ". Non_pertinente	Smart Living & Energy,	Smart Manufacturing
"Development and validation of an automated process for the production and assembly of synthetic DNA using Smart Algorithms and Robotics"	The consolidation project presented by the innovative start-up Officinae Bio has the aim of developing a platform for the design and production of synthetic DNA intended for companies in the Cell & Gene Therapy sector (CGT). The MVP made by the company, already complete from a software point of view, will be implemented through the robotic automation of the hardware system. In particular, the prototype will be enriched through the addition of new features that will allow to automate, optimize and integrate the different phases of the process, in order to structure a large -scale production. Thanks to the superior technical skills of one's DNA, combined with a solid production capacity, workshop Bio can not only offer the market a performing and safe product with reduced times and costs, but also to become a point of reference for the development and innovation of The entire CGT sector. non_pertinent	Smart Manufacturing, Space economy	Smart Health

Conclusion

The automatic classifier models have been developed and implemented. All the projects, without previous classification according to the S3 of the Veneto Region, that are inserted in the database of the dashboard have been analysed and classified by these models.

The applications of these classifiers are multiple: they can be used, for example, to predict any new given text, to evaluate benchmark regions' specialisations, or to classify any other type of textual documents (such as scientific publications and any new funding programme's projects).

These models can be improved, through human iterations, and updated, for example by adding new categories and possibilities of classifications. The possibility or other developments and improvements in this sense will be discussed and evaluated in further collaborations.