A machine-learning approach for a CRIS research outputs' SDG classifications

Authors:

António Lopes, Catarina Roseta-Palma, Ana Simaens

Iscte – Instituto Universitário de Lisboa, Lisbon, Portugal

EuroCRIS Autumn 2023 Strategic Membership Meeting, Pamplona, Spain

Presentation Proposal

Introduction

The Sustainable Development Goals (SDGs) were defined by the United Nations in 2015 to provide goalposts for humanity's ambition to move towards a better planet, a prosperous economy, and an inclusive society by 2030. We stand roughly at the midpoint of the SDG implementation period, a good moment to take stock of developments. Higher education institutions (HEIs) play a fundamental role in the creation of knowledge and its dissemination, so they are crucial levers to ensure that the SDGs reach a wider audience.

Accordingly, various studies have summarized the contributions of HEIs to the various SDGs in terms of their strategy (Leal Filho et al., 2023), education (Leal Filho et al., 2019), sustainability reporting in the sector (De la Poza et al., 2021) and, especially, research (Agnew et al., 2020). Many individual HEIs have committed to Agenda 2030 and wish to assess their own contributions to the SDG, yet not all have the resources to apply such methodologies themselves: many lack an adequate database of publications with SDG relevance, since manual assessments are time-consuming and might not reflect widely accepted categories.

Machine-learning (ML) can be a valuable tool in this task of automatically classifying scientific publications as to their contribution to the SDGs (Angin et al., 2022; Morales-Hernández et al., 2022), allowing HEIs to monitor their own contributions and appraise their impact as well as improving communication and increasing the engagement of the academic community. Current Research Information Systems (CRIS) are ideal candidates for deploying this kind of approach because they can combine the availability of research outputs and external communication features with internal machine-learning models to help researchers choose the most accurate SDGs for which their research output contributes to.

This presentation takes stock of the methodologies that have been applied to the assessment of research outputs as they relate to the SDGs, in our Institution's CRIS,

Ciência-IUL[1]. In particular, we focus on the machine-learning-based approach that was employed in the CRIS to help researchers choose the right SDGs to be associated with their research outputs (including publications and projects).

Our Approach

This work started in July 2018, when all researchers were invited to start classifying their research outputs in terms of their contributions to the SDGs. This global assessment allows depicting a clear picture of the entire University's contributions to the SDGs. Figure 1 shows a screenshot of the CRIS' public page dedicated to the Sustainable Development Goals.

Figure 1 - Iscte's publications and projects and their contributions to the SDGs as seen in the CRIS

The page depicts the number of publications and projects associated with each SDG, but it is possible to browse a specific SDG and retrieve a detailed list of all publications and projects that contribute to that SDG. Figure 2 shows the example of the list of publications and projects that contribute to SDG 4.

Manually classifying all the research outputs in the CRIS can be very time-consuming. To facilitate this task, we started working on building an automated approach that could help researchers by suggesting which SDGs are the most suitable for their respective outputs. Figure 3 depicts the timeline of our entire approach. In June 2021, we gathered all the data collected so far regarding manual classifications of research outputs contributions to SDGs performed by the University's researchers (a dataset with 9665 records) and started testing, training and evaluating different machine learning algorithms. In September 2021, we integrated the developed model within the CRIS and by January 2022, researchers were starting to get automated suggestions when adding new research outputs in the system.

[1] http://ciencia.iscte-iul.pt/

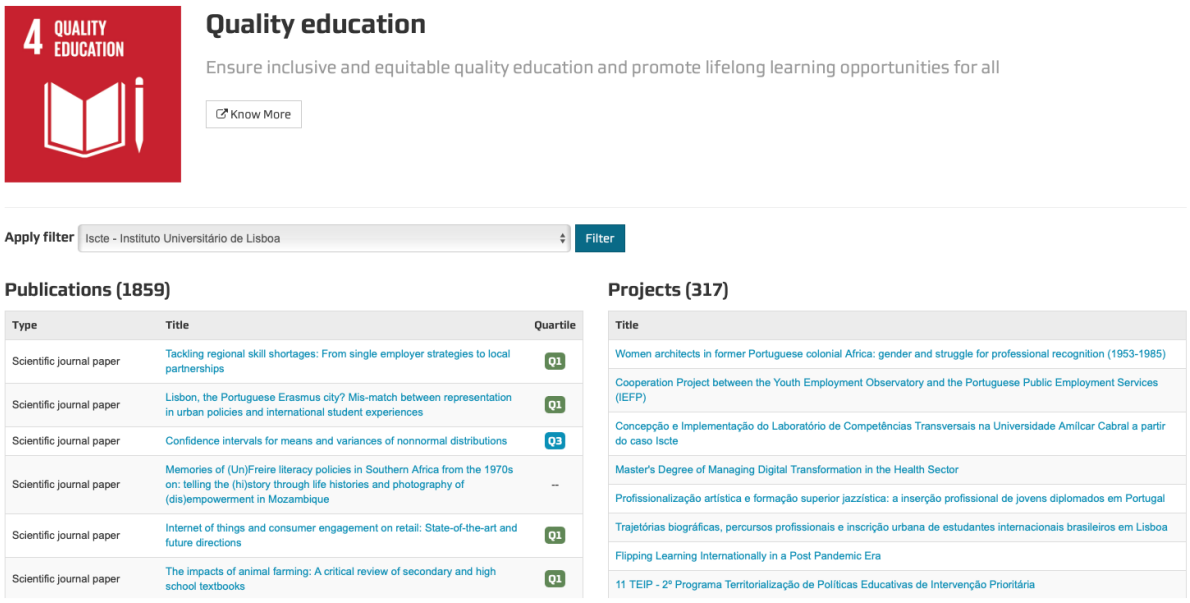

Figure 2 - Iscte's publications and projects that contribute to SDG4 as seen in the CRIS

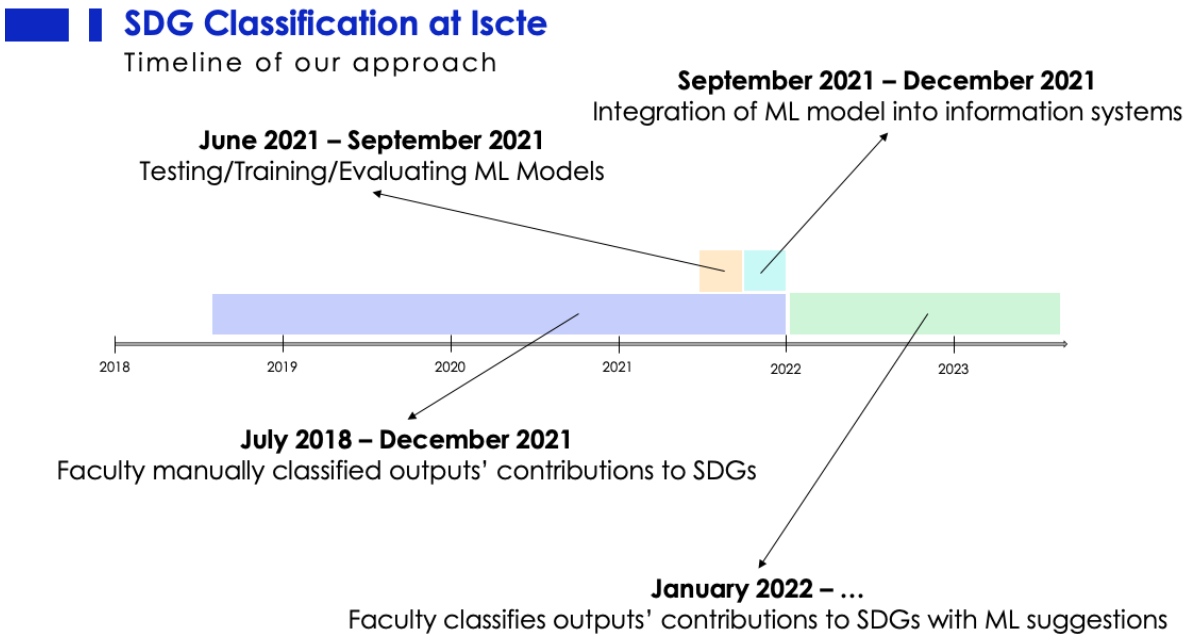

Figure 3 - Timeline of our approach in developing and integrating a ML Model into the CRIS

The training process is an intensive one, where we provide a list of annotated text records to our algorithm, in which for each record we indicate the SDGs for which the record contributes to. The whole idea of this process is to provide the algorithm with enough information so that it can capture the relationship between the words in the text records and a particular SDG. Afterwards, this allows us to provide the algorithm with a completely

new record (never before seen) and ask it to predict which SDGs should be associated with this new record. And the algorithm will be able to provide that prediction, with a certain level of confidence. Figure 4 shows how the suggestions appear in the CRIS.

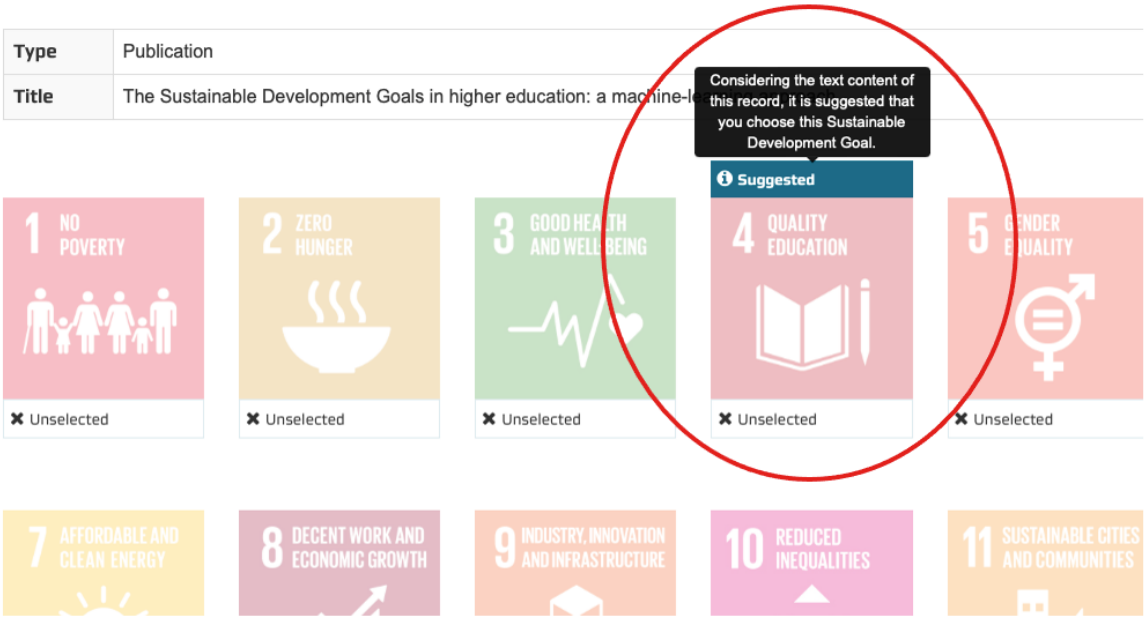

Figure 4 - Suggestions for SDG as shown in the CRIS

We tested the following algorithms with a dataset of 9665 annotated records:

- Gaussian Naive Bayes (**GNB**)
- Multinomial Naive Bayes (**MNB**)
- Linear Support Vector Classification (**SVC**)
- Logistic Regression (**LR**)

We used the following metrics to evaluate the performance of each algorithm:

- **ACCURACY** – ratio of correctly classified samples
- **RECALL** - ability to find all the positive samples
- **PRECISION** – ability to identify all the positive samples without accidentally marking too many negative samples as positive
- **F-MEASURE** - harmonic mean of the Precision and Recall

Figure 5 shows the accuracy metric values for each algorithm and for each SDG. The SVC algorithm was a clear outlier in most SDGs, so that is the one we decided to use in the deployment of the suggestions feature in the University's CRIS.

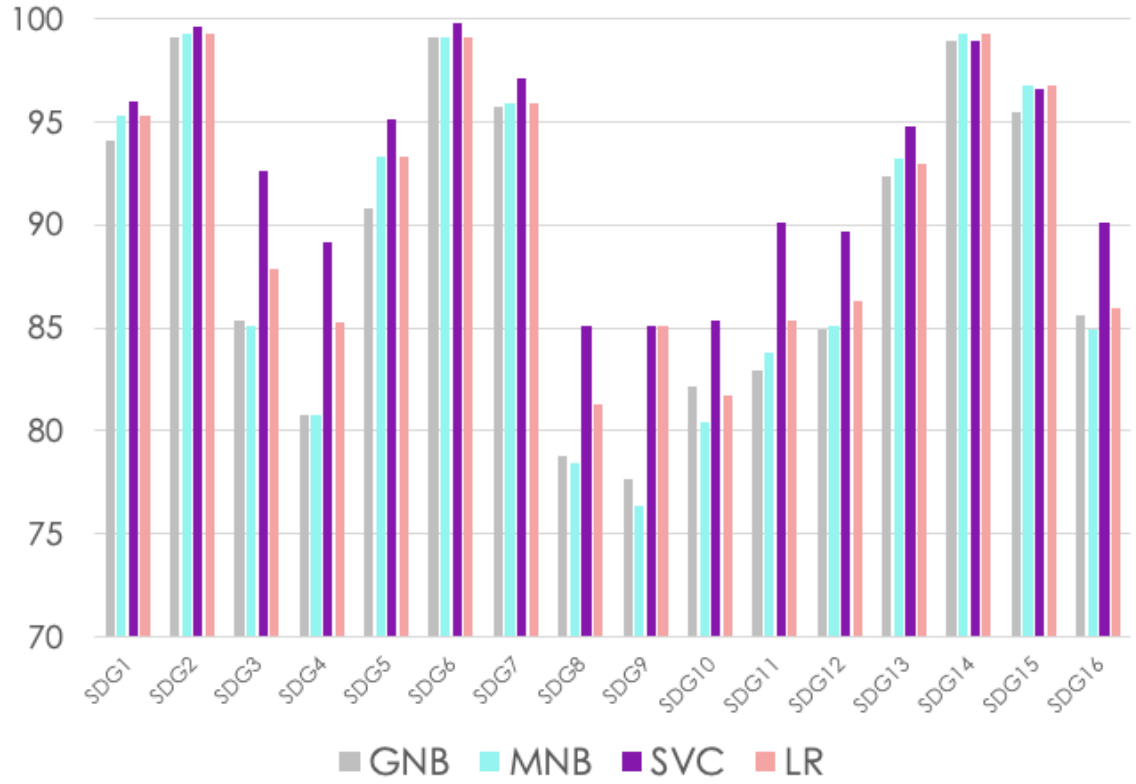

Figure 5 - Accuracy performance for each of the algorithms in each of the SDGs

When comparing the ML model's overall performance between its choices and the choices of the researchers, the model showed an 87% accuracy on average. This shows that the ML model is successful in suggesting SDGs that should be associated with outputs. We believe this value can be further improved by leveraging other techniques that we intend to explore in future work.

References

Agnew, K., Francescon, D., Martin, R., Rhannam, M., Schemm, Y., Balisciano, M., Bayazit, K., Bos, C., Erkal, E., & Falk-Krzesinski, H. J. (2020). "The Power of Data to Advance the SDGs: Mapping research for the Sustainable Development Goals." https://www.elsevier.com/__data/assets/pdf_file/0004/1058179/Elsevier-SDG-Report-2020.pdf

Angin, M., Taşdemir, B., Yılmaz, C. A., Demiralp, G., Atay, M., Angin, P., & Dikmener, G. (2022). A RoBERTa Approach for Automated Processing of Sustainability Reports. In Sustainability (Vol. 14, Issue 23). https://doi.org/10.3390/su142316139

De la Poza, E., Merello, P., Barberá, A., & Celani, A. (2021). Universities' Reporting on SDGs: Using THE Impact Rankings to Model and Measure Their Contribution to Sustainability. In Sustainability (Vol. 13, Issue 4). https://doi.org/10.3390/su13042038

Leal Filho, W., Shiel, C., Paço, A., Mifsud, M., Ávila, L. V., Brandli, L. L., Molthan-Hill, P., Pace, P., Azeiteiro, U. M., Vargas, V. R., & Caeiro, S. (2019). Sustainable Development Goals and sustainability teaching at universities: Falling behind or getting ahead of the pack? Journal of Cleaner Production, 232, 285–294. https://doi.org/10.1016/J.JCLEPRO.2019.05.309

Leal Filho, W., Simaens, A., Paço, A., Hernandez-Diaz, P. M., Vasconcelos, C. R. P., Fritzen, B., & Mac-Lean, C. (2023). Integrating the Sustainable Development Goals into the strategy of higher education institutions. International Journal of Sustainable Development & World Ecology, 1–12. https://doi.org/10.1080/13504509.2023.2167884