**1. Project Name**

Machine Learning Research for SDG Classifier

**2. Overview**

This Working Group will contribute to the development of the OSDG[[1]](#footnote-2),[[2]](#footnote-3), which is an open-source tool that assigns labels to document contents based on the SDGs resemblance. The workstreams are part of the overall PPMI/OSDG-SDG AI LAB Partnership Concept for “Advancing AI, ML, and NLP Research for the SDG Analysis”

Working Group 3 will specifically focus on engaging a research on supervised and unsupervised Machine Learning (ML) methods. The main objective is to classify textual data to Sustainable Development Goals (SDGs) based on an annotated dataset and ontology created from Working Group 1 and 2. This research stream will examine the general approaches optimized for NLP Text Classification tasks, as well as possible ML algorithm adaptation to the OSDG tools. Through a comprehensive examination, performance of the different ML models will be compared for the given task and based on the result, the new model(s) will be prepared for implementation. Main deliverables are the trained and evaluated machine learning model(s), the research report for the ML models, including data exploration, description and evaluation and testing metrices. Moreover, WG 3 is encouraged to add functionalities that allow to identify why the model makes a certain decision.

**3. Background**

There is a strong need and demand from the UN System, public institutions, and private sector for classifying government publications, policy briefs, academic literature and CSR reports according to their relevance to the Sustainable Development Goals (SDGs). It is well understood that the SDGs play major role in the strategic objectives of various entities. However, linking projects and activities to the SDGs has not always been straightforward or possible with the existing methodologies.

Machine Learning (ML) methods, in combination with an ontology, is the current approach employed for classifying textual data under the PPMI/OSDG-SDG AI LAB Partnership. As the size of ontology is increasing, the performance of ML model also has the potential to improve through a training and a feature engineering process. The contributions from this working groups will be integrated into the OSDG, open-source tool that assigns labels to content based on SDGs.

Supervised ML approaches work by taking a set of input data together with the output labels and using a dedicated algorithm to find the best way to match the input data with the outputs. Unsupervised ML does not aim to predict a specific label, but rather to find the underlying structure and clusters in the data. An example of this is the [￼[[3]](#footnote-4)](https://sdg-pathfinder.org/)￼the OECD that utilizes topic modelling on the pooled corpus of documents.

OSDG tool uses a hybrid approach combining Ontology and ML. Individual ontology items are derived from the ontology-building process and method of measuring the most significant features for each SDG label comes from the Supervised ML models. Beta coefficients from the logistic regression models are currently used in place. In the next stage of the project, the plan is to develop a supervised ML model using said corpus as a training set.

**4. Key Objectives & Research Questions**

The key objective of this workstream is to investigate comprehensive ML methodologies and to optimize the task performance of deriving keywords and assigning labels to a text paragraph based on the SDGs. Foundations of the work has been established by the Team 4 of SDG AI Lab and the new Working Group will build upon the progress made. For the primary objective, ML workflows will be comprehensively examined and compared. The plan is to develop a supervised ML model using either or both the newly created ontology from Working Group 1 and the new corpus from Working Group 2 as a training set. As the secondary objective, this workstream will test different ML models to identifying significant/important features from a document. This phase also involves testing different classification approaches, namely multi-class and multi-label. Depending on the annotation level from working group 2 and the agreed objectives, also hierarchical classification approaches that dive into the target/indicator level may be considered. Through this quantitative comparison in performance, it is expected to identify the approach that delivers the best results for the task at hand.

Throughout the work, the research questions that need to be answered are the following:

RQ1: In comparison with the known supervised (and unsupervised) ML methods, which approach does suit the best for categorizing text-based information to the SDGs?

RQ2: How accurately can we classify text documents to one or more SDGs or a deeper level of granularity?

RQ3: How can we ensure the quality of the trained models? How can we use domain expert input to evaluate and improve the models after training?

RQ4: Based on the findings, what are the suggestions for new and better rules for training data annotation and integration to improve performance of the models?

**5. Workflows Ideas and Suggestions**

The project will start with a review of the list of supervised and unsupervised Machine Learning models, and then initiate following the problem-solving patterns to create a result step by step. A potential aspect to focus amongst various approaches is to build a multi-label supervised learning classifier, a multi-class classifier or enhance performance of the existing OSDG tool with the annotated data. Subsequently, the annotated datasets from WG 2 are processed and most suitable features, models are engineered, evaluated, extensively tested and compared. While preparing for reports and presentation, which are to be completed by the end of the journey, it is expected to address the shortcomings and challenges encountered on the way. SDG AI Lab will revisit the outcomes of the work streams following an agile approach that will be implemented through task-based, timeboxed iterations, and rely on asynchronous and remote cooperation.

Note that these suggestions are first and foremost guidelines. Please consider the objectives as the main focus, but may choose other methodologies and models that deems more suitable or accurate for the tasks.

**6. Dataset and File Descriptions**

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Description | Folder Name | Link |
| 0. | SDG Ontology compiled by Dr Nuria B. Puig and E. Mauleon | 0\_PuigOntology | [Dataset](https://figshare.com/articles/SDG_ontology/11106113/1) |
| 1. | Mapping from "FP7-4-SD" Project (edited VS and LP) | 1\_FP7-4-SD\_edited | [Link to Project website](https://www.fp7-4-sd.eu/) |
| 2. | Concepts UN Linked SDG tool extracted from academic publications | 2\_LinkedSDG\_Concepts | [Link to LinkedSGS Tool](http://linkedsdg.apps.officialstatistics.org/#/) |
| 3. | Concepts extracted from SDG Pathfinder documents extracted via ML | 3\_SDGPathfiner\_DocumentConcepts | [Document Colletion](https://sdg-pathfinder.org/) ; [Modelling Description](https://ppmi.lt/) |
| 4. | Keywords from SDG Pathfinder indicated by the SDG Pathfinder tool itself | 4\_SDGPathfinder\_Keywords | [SDG Pathfinder](https://sdg-pathfinder.org/) |
| 5. | Concepts UN Linked SDG tool extracted from Administrative Documents | 5\_LinkedSDG\_DocumentExtracts | [Link to LinkedSGS Tool](http://linkedsdg.apps.officialstatistics.org/#/) |
| 6. | Terms by Indicator from SDGIO Ontology | 6\_SDGIO\_Terms | [Link to SDGIO GitHub](https://github.com/SDG-InterfaceOntology/sdgio) |
| 7. | Concepts linked to SDGs from EC Policy Documents | 7\_EC\_Policy\_Doc\_Terms | Skrynnyk & Stanciauskas ( 2020 upcoming ) |
| 9. | Keywords from "Science4SDGs" project | 9\_SIRIS\_Science4SDGs | [Link to "Science4SDGs" project](http://science4sdgs.sirisacademic.com/) |

**7. Deliverables**

* A research report and a presentation covering the ML methods for document classification task, including quantitative performance evaluation to compare the results of various models.
* One or more trained machine learning models that allow to infer unseen documents or texts reliably and accurately to the SDGs on the desired level of granularity.

1. Official Website <https://www.osdg.ai/> [↑](#footnote-ref-2)
2. Pukelis, L., Puig, N. B., Skrynik, M., & Stanciauskas, V. (2020). OSDG--Open-Source Approach to Classify Text Data by UN Sustainable Development Goals (SDGs). *arXiv preprint arXiv:2005.14569*. [↑](#footnote-ref-3)
3. Official Website <https://sdg-pathfinder.org/> [↑](#footnote-ref-4)