Automatic classification of multilabel texts related to Sustainable Development Goals (SDGs)

Jade Guisiano, Raja Chiky

Automatic classification of multilabel texts related to Sustainable Development Goals (SDGs)

Jade Guisiano*, Raja Chiky**

*United Nations Environment Program
jade-guisiano@outlook.fr,
https://www.unenvironment.org
**Institut Supérieur d'Electronique de Paris
raja.chiky@isep.fr
https://www.isep.fr

Abstract. The Sustainable Development Goals (SDGs) are the guiding line to achieve a better and more sustainable future for all. They tackle the global challenges we face, including poverty, inequality, climate change, environmental degradation, peace and justice. The OnePlanet network - an open partnership for sustainable development - provides a platform where all countries, including all relevant stakeholders and organizations, are invited to join and actively engage. Thus, the latter submit daily a large number of descriptions of innovative projects that may be linked to one or more SDGs. For experts, the task of linking all the texts submitted to the SDGs they deal with is very time consuming, which is why the need to automate this process is very important. In this context, we propose to solve this problem with a multi-lablel classification of texts using BERT (Bidirectional Encoder Representations from Transformers). We first present the key steps for building our database, then the multi-label classification phase using BERT and finally we will present and discuss the obtained results.

Context and objective

One Planet Network [1] focuses its programs on Sustainable Consumption and Production corresponding to the Sustainable Development Goals number 12 and they collect a great number of project propositions linked to their activities. The Sustainable Development Goals (SDGs), also known as the Global Goals, were adopted by all United Nations Member States in 2015 as a universal call to action to end poverty, protect the planet and ensure that all people enjoy peace and prosperity by 2030 [2] . SDGs are interdependent, and it is not uncommon for one SDG to be linked to another. The task to link this huge amount of submitted text to SDGs it deals with is very time-consuming for experts, that is why the One Planet Network identified the need to automate this processing. The main objective of the analysis is to to establish the degree to which a project proposal is linked to their activity (i.e. to SDG12) but also to the

1. https://www.oneplanetsummit.fr/
2. https://www.undp.org/content/undp/en/home/sustainable-development-goals.html

remaining 16 SDGs. In fact, the objectives categorized by the SDGs being interconnected, a project description is frequently related to many SDGs.

Existing solution and its limitations

First, an initial state of the art has been established and we identified a method (Pincet et al., 2019) from the OECD (Organisation for Economic Co-operation and Development). This method consists of implementing an automatic text classification using Universal Language Model Fine-tuning (ULMFiT) (Howard and Ruder, 2018) , it allows us to determine to which (and only one) of the 17 SDGs a text is linked. However, this method is not adapted to our problem for 3 main reasons:

— This method allows to link a text only to a single SDG.
— The degree of correspondence of a text to a SDG is not quantified.
— The training of the algorithm requires a labeled text set (Labeled text is a designation for pieces of text that have been tagged with one or more labels identifying certain properties or characteristic, in our case labels are SDGs) which means that for each input text the SDG to which it is linked must be established downstream by an expert (which can be very time consuming).

For these reasons we decided to carry out further research in order to build a method that permits to link a text to several SDGs but also permits to quantify its degree of belonging to the different SDGs.

Database

As we need a labeled database, we use information extracted from the United Nations website and more especially the 2030 Agenda for Sustainable Development's description [3]. From this information we extract each available target (which constitutes the input text) from each of the 17 SDGs (which constitute our labels). We construct a CSV file (composed of 169 lines) with a column containing texts (targets) ID, the text itself and 17 columns for the 17 SDGs. To link a text with its SDG number, we choose the one-hot-encoding format (which consist to assign the value 1 in a SDG column at the line where a considered text belongs to the SDG).

But as this stage the dataset size is still too small and it is composed of only UN institutional text style. That is why we decided to increase the size of our dataset and also diversify it with other texts. For that, we realized 3 additional steps :

— Synonym's definition : by simple statistic analysis we could obtain the most occurrent words for each SDG texts, we then conserve the top 3 of them and applied NLTK [4] Python library which helps us to build a list of synonyms for each keyword. Finally, we use the Wikipedia API [5] for each synonym to extract its definition. Then we add the definition for each SDG.

3. https://sustainabledevelopment.un.org/post2015/ transformingourworld
4. https://www.nltk.org
5. https://pypi.org/project/Wikipedia-API/

— UN expert from One Planet also provide a folder containing 1 to 3 PDF text about each SDG. We then extract the text from each PDF, clean it and add it to our database.
— Finally we decided to use Markovify [6] (in order to generate 300 additional texts for each SDG) which works as follow: In the first step, it learns what kind of text is associated to each SDG and then it creates similar texts (treating about the same subject), then it generates strings of words that are probably linked. The generation is entirely random and based on the probabilities of associations between each word.

The final dataset labeled we obtained is "res.csv" which contains 6017 lines of text and its associated SDG.

Multi-label Text Classification

As our goal is to quantify the degree to which a text belongs to all 17 SDGs, to achieve it we opted for a Multi-labeled Text classification algorihtm, which makes it possible to link a text not only to one label (SDG) but to several labels. We also decided to include in our method a recent powerful algorithm adapted to treat NLP tasks, this algorithm is named BERT (Delvin et al., 2018).

BERT

BERT [7]: Bidirectional Encoder Representations from Transformers is a neural network architecture designed by Google researchers that's totally revolutionized and surpassed the state-of-the-art for NLP tasks such as text classification, translation, summarization, and question answering. Indeed, unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers. Concerning the data processing part, we first split our dataset "res.csv" in train and test, the splitting ratio is 80% for train and 20% for the test. Then, we need to transform our data into a format which BERT can understand. For that we create the class "InputExample's" using the constructor provided in the BERT library with variable "text_a" containing the text we want to classify and "label" is the labels corresponding to our texts. In a second step we need to pre-process our text data so that it matches the BERT's requirements (text in lower case, tokenization, etc.)

Then, we load BERT module and create a single new layer that will be trained to adapt BERT to our Multi-label classification task. The training phase has the goal to train the model to learn the main characteristics (words and their context) of the texts contained in our 17 classes (17 SDGs). Then the model keeps its training parameters to be applied to texts where their class is not given to the algorithm (But we keep them apart.), the algorithm with its parameters must be able to classify them correctly and once the classification is established we compare the class to which the algorithm classifies the text to the one to which it is really assigned. Then, Our goal being to quantify the belonging of a text to a SDG, we set the prediction output in probability thanks to the use of a Logit function. We also include Dropout

6. https://datascienceplus.com/natural-language-generation-with-markovify-in-python/
7. https://github.com/google-research/bert

to prevent overfitting [8] And finally we compute the loss between predicted and real label which permits to assess the accuracy of our model which has an accuracy of 94,21%.

Results

In order to test our trained model we take 3 texts extracted from the SDG pathfinder [9]. Pathfinder provides text which are associated to the SDG they treat thanks to a machine learning method. This first result comparison permit us to check our method performance compared to an other text classification algorithm (In other words, we want to check if our algorithm result is close to the existing classification from the pathfinder website).

— Text 1 : (SDG 14 - Life Below Water)

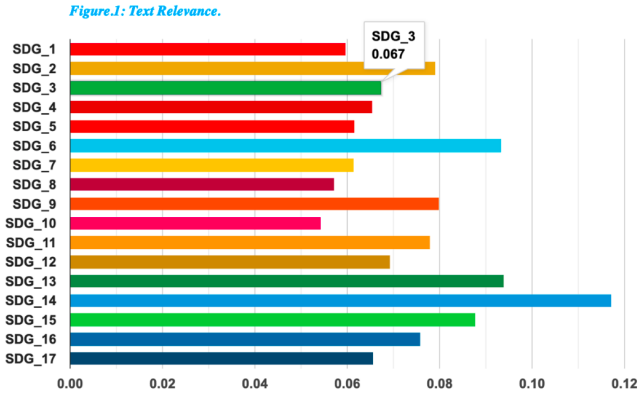

FIG. 1 – *Results for Text 1.*

We clearly see on the figure 1 that our method result present a spike for the SDG14 which mean that our algorithm detect that the text mainly treats about SDG14. We also see that other SDG subjects are detected by our algorithm, as we do not know the exact machine learning algorithm uses by Pathfinder it is difficult to make a conclusion on this other detected SDG. May be an expert can determine if this detected SDG (6 and 13) is relevant for this text.

— Text 2 : (SDG 4 - Quality Education)

8. Overfitting occurs when the algorithm over-learns (overfit.)in other words, when it learns from data but also from patterns (diagrams, structures) which are not related to the problem, such as noise, thus degrading the performance of the algorithm.

9. https://sdg-pathfinder.org

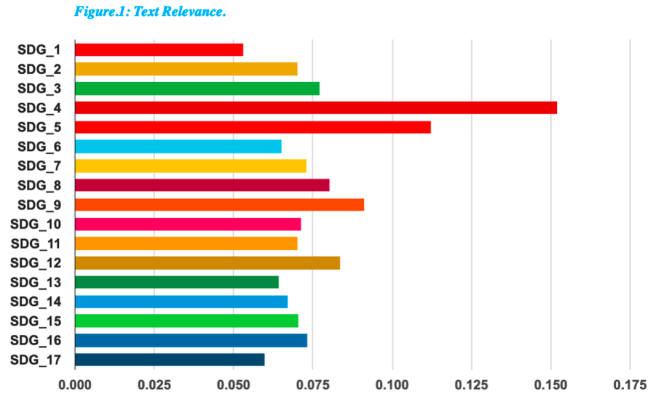

FIG. 2 – *Results for Text 2.*

The figure 2 show that our method result present a spike for the SDG4 so that our algorithm find the good SDG subject.

— Text 3 : (SDG 13 - Climate Action)

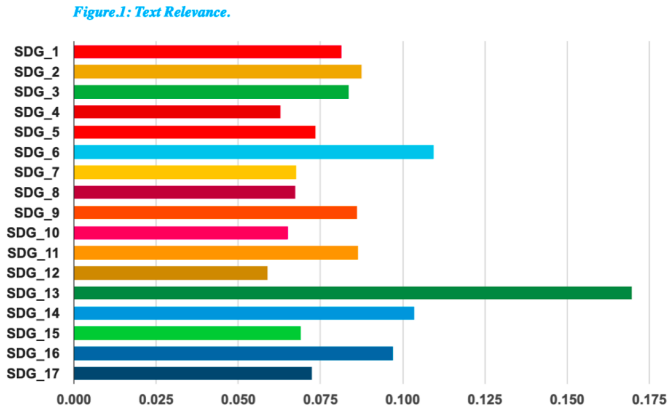

FIG. 3 – *Results for Text 3.*

We see on the figure 3 that our method result indeed present a,clear peak for the SDG14.

We also compare our method with manually expert text classification available on the website IISD which provide news article and indicates which SDGs they belong to. We took an article treating about intituled "COVID-19: What Happens When Summer Heatwaves Strike?" [10] which treats about SDG 3 (good health and well-being), SDG 11 (sustainable cities and communities), SDG 13 (climate action) (odre of importance in the text). Our method provides the following results :

10. http://sdg.iisd.org/commentary/guest-articles/covid-19-what-happens-when-summer-heatwaves-strike/

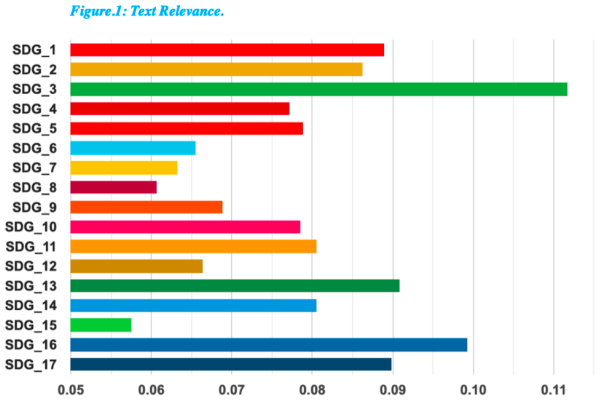

FIG. 4 – *Results for Text 4.*

We see in the figure 4 that our results present a clear spike for the SDG 3, then SDG 11 and finally SDG 13. This is the exact SDG (even in the good order of importance) initially labelled by experts.

Conclusion

We proposed in this paper a solution for Multilabel text classification in order to link a text to the SDGs it deals with. Our solution is based on a recent and efficient Natural Language Processing (NLP) method called BERT. The carried experimentation show the efficiency of our approach that gives an accuracy of 94,21% during the training phase. We then tested our model on real cases with texts providing from new sources. One source contained text already linked to an SDG — this annotation was realized by other automated classification algorithm — and our algorithm was able to identify the same SDG. We also test our method on an other source which provide text classified by experts and our method was also able to find the same result. To go further, it would be interesting to enrich the database by using various and heterogeneous sources related to SDGs with different vocabularies. In this next point, if text are well chosen, our method performance could be event better than actual one.

References

Delvin, J., M. Chang, K. Lee, and K. Toutanova (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

Howard, J. and S. Ruder (2018). *Universal Language Model Fine-tuning for Text Classification*.

Pincet, A., S. Okabe, and M. Pawelczyk (2019). Linking aid to the sustainable development goals – a machine learning approach. *OECD Development Co-operation Working Papers 52*.

Résumé

Les Objectifs de Développement Durable (ODD) sont la ligne conductrice pour parvenir à un avenir meilleur et plus durable pour tous. Ils s'attaquent aux défis mondiaux auxquels nous sommes confrontés, notamment la pauvreté, les inégalités, le changement climatique, la dégradation de l'environnement, la paix et la justice. Le réseau OnePlanet - un partenariat ouvert pour le développement durable - fournit une plate-forme où tous les pays y compris toutes les parties prenantes et organisations concernées, sont invités à se joindre et à s'engager activement. Ainsi, ces derniers déposent quotidiennement un grand nombre de description de projets innovants pouvant être liés à un ou plusieurs ODD. Pour les experts, la tâche de lier tous les textes soumis aux ODD dont ils traitent est très chronophage, c'est pourquoi le besoin d'automatiser ce processus est très important. Dans ce contexte, nous proposons de résoudre cette problématique avec une classification multi-lablel de textes à l'aide de BERT (Bidirectional Encoder Representations from Transformers). Nous présentons dans un premier temps les étapes clés de la constitution de notre base de données, puis la phase de classification multi-label à l'aide de BERT et enfin nous présenterons et discuterons les résultats obtenus.