Software Impacts 14 (2022) 100427

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

EnvBert: An NLP model for Environmental Due Diligence data classification

Afreen Aman [a,*], Deepak John Reji [b]

[a] Ernst & Young, India
[b] Environmental Resources Management, India

ARTICLE INFO

Keywords:
Natural Language Processing
Environmental Due Diligence
DistilBERT
EnvBert
Hugging Face

ABSTRACT

EnvBert is a Natural Language Processing (NLP) package built on top of DistilBERT, focused on Environmental Due Diligence (EDD). It comprises of four functions: text classification, relevancy detection, ranking and fine-tuning. Environmental dataset was used for fine-tuning DistilBERT's performance to develop EDD model. It is hosted as an inference Application Programming Interface (API) on Hugging Face Hub. The EDD model with custom vector representation constitutes EnvBert, it can be installed using the package manager pip.

Code metadata

Current code version	v1.0.6
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2022-207
Permanent link to reproducible capsule	https://codeocean.com/capsule/6685103/tree/v1
Legal code license	MIT License
Code versioning system used	git
Software code languages, tools and services used	python, PyPI
Compilation requirements, operating environments and dependencies	Python 3.6 and above
If available, link to developer documentation/manual	https://github.com/dreji18/Environmental-Due-Diligence/blob/main/README.md
Support email for questions	afreen.aman@gds.ey.com, deepak.reji@erm.com

EnvBert

Environmental Due Diligence (EDD) is a process of collecting, assessing and evaluating data related to environmental conditions to support industrial site concessions, property acquisition, and corporate expansions or mergers. The aim of EDD is to ensure compliance with environmental law, detect environmental liabilities and estimate the costs associated with it, reduce reputational risks and risks of legal litigations [1]. This process involves reviewing the records maintained by regulatory bodies, collating the data, interviews with the occupants of the land and neighbours, and preparation of a site assessment report. The whole process is textual data focused, thereby serves as an apt use case for NLP.

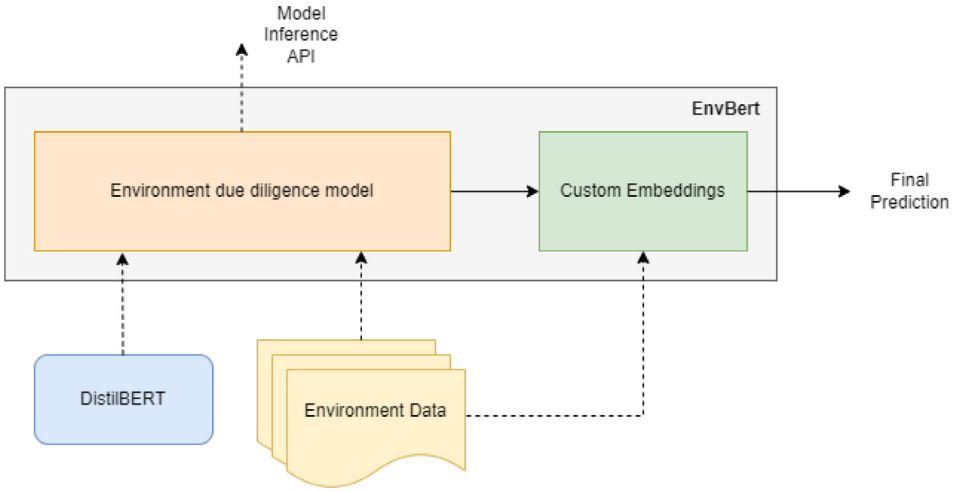
We built an easy-to-use Python library, "EnvBert" with the EDD model and custom curated vectors to identify essential environmental data as a part of environment site assessments in due diligence. EnvBert has the following capabilities:

Feature	Output
EDD Prediction	Categorizes the Environmental data under different classes
Relevancy	Classifies whether it is relevant or not for the Environmental domain
Ranking	Returns Relevance probability against the predicted classes
Fine-tuning	Allows users to train the model for their Environmental dataset

EnvBert is a pipeline solution which has an Environmental Due Diligence classification model followed by custom word vectors which

DOI of original article: https://doi.org/10.1016/j.dib.2022.108579.
The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.
* Corresponding author.
E-mail addresses: afreen.aman@gds.ey.com (A. Aman), deepak.reji@erm.com (D.J. Reji).

https://doi.org/10.1016/j.simpa.2022.100427
Received 16 September 2022; Received in revised form 26 September 2022; Accepted 28 September 2022



predicts the probability of a given sentence/article against a set of categories [2]. The classification model is trained on customized environmental dataset to detect contamination and remediation activities (both prevailing as well as planned) as a part of site assessment process. It can classify the data into the following categories: remediation standards, extent of contamination, depth to water, groundwater-surface water interaction, groundwater velocity, geology, contaminated media, remediation activities, remediation goals, source of contamination, contaminants and not relevant.

This model was built on top of distilbert-base-uncased model and trained for 10 epochs using TensorFlow framework with a batch size of 16, a learning rate of 5e−5, and a maximum sequence length of 512. We used Code Carbon to track the carbon emission generated while training the model. It was found to be 0.1069 kg. This model is hosted on Hugging Face hub, and it can be inferenced using transformers library or as an API endpoint provided by Hugging Face Accelerated Inference [3].

The output from this language model is passed on to a custom word embeddings layer which boosts the performance of model prediction.

EnvBert python package can be installed using the package manager pip:

```
pip install EnvBert
```

Train accuracy	*Validation accuracy*	*Train loss*	*Test loss*
78.2%	*71%*	*0.42*	*0.92*

EnvBert can be used to predict the environment categories as well as fine-tune on custom dataset of user's choice. The function: 'envbert_predict' returns the predicted class along with the probability. The 'finetune' component allows training based on the training configuration with custom environment data and labels. The model is trained internally with TensorFlow framework, and it saves the model files and tokenizer in a directory of choice. The fine-tuning happens over the environment due diligence model which is the first phase of EnvBert pipeline. Once the model is saved it can be loaded using finetune_predict component.

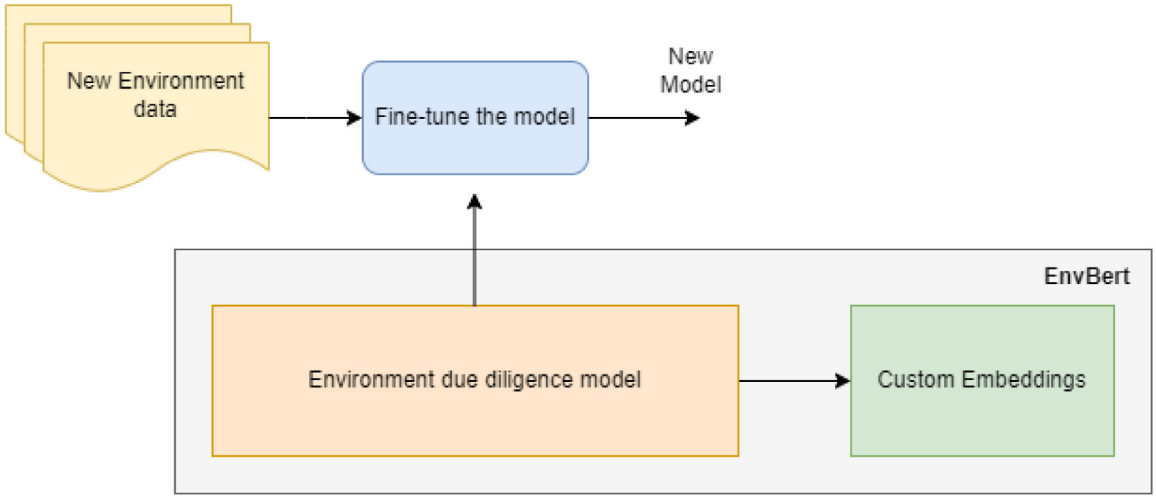
The major benefit of using EnvBert is that with a single line of code, users would be able to train their environment specific language model and use it for their custom use case.

Software Impacts

Deep neural networks have pushed the frontier in NLP by excelling at learning from labelled data and achieving state-of-the-art results on a wide range of NLP tasks. Pre-trained models like ELMo [4] and BERT [5] have outperformed static word embeddings on many NLP tasks. However, it is challenging to get the desired results from NLP models in unfamiliar conditions. This pushes forward the need for domain adaptation in models. Traditionally, domain in NLP, is referred to a coherent type of corpus, which is generally predetermined by a given dataset. We trained our model on EDD data which was handpicked from documents released by regulatory bodies for 11 environmental categories and the data that did not fit into these categories was labelled as 'Not Relevant'. Therefore, the model categorizes the data into 12 labels. The data used to train the model can be found at Mendeley [6]. This dataset can be used by users to train NER models, recognize and extract entities, train QnA models, perform topic modelling etc.

Fine-tuning is one of the methods to achieve high performance across NLP tasks wherein the transformer-based model is trained with a small amount of labelled data. It involves using the pre-trained model weights and training a new layer on domain data [7]. Our classification model is fine-tuned on custom curated environmental dataset to detect various classes of environmental due diligence to help the users in the preparation of environmental due diligence reports and gives the user liberty to train more classes by using minimal amount of training data.

Environmental remediation/ restoration projects, lifecycle site evaluation, environmental construction services require assessment of past and current environmental condition of the contaminated area before developing remediation strategies. Our model can ease the process of data collection and assessment pertaining to a contaminated site. It prioritizes the information for each category by using embedding's



similarity score. This model can also help in identification of emerging contaminants and assess information related to cleanup activities.

As a future improvement, this model can be made more generalized and trained for more categories. The model's focus is on contaminated site assessment classes, the scope can be expanded by developing more models specifically targeted to suit the environmental data requirements of Mergers and Acquisitions, property acquisitions, company expansions etc. The model's scope can be extended by using it to identify and classify the EDD data from documents in different languages with the help of a language translation service in the encoder and decoder section of the model.

'Environmental Due Diligence Data: A novel corpus for training environmental domain NLP models', an article in the Data in Brief Journal describes the dataset that was used to train the EnvBert Model and its applicability for a variety of NLP tasks [8].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A.R.D. Mareddy, A. Shah, N. Davergave, Environmental Impact Assessment: Theory and Practice, Butterworth-Heinemann, 2017.
[2] A. Aman, D.J. Reji, Envbert, 2022, https://pypi.org/project/EnvBert/.
[3] D.J. Reji, A. Aman, D4data/environmental-due-diligence-model, 2022, d4data/environmental-due-diligence-model Hugging Face.
[4] M.E. Peters, M. Neumann, L. Zettlemoyer, W.T. Yih, Dissecting contextual word embeddings: Architecture and representation, 2018, http://dx.doi.org/10.48550/arXiv.1808.08949, arXiv e-prints, arXiv-1808.
[5] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, http://dx.doi.org/10.48550/arXiv.1810.04805, arXiv e-prints, arXiv-1810.
[6] A. Aman, D.J. Reji, Environmental Due Diligence Data, Vol. V4, Mendeley Data, 2022, http://dx.doi.org/10.17632/tx6vmd4g9p.4.
[7] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, N. Smith, Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020, http://dx.doi.org/10.48550/arXiv.2002.06305, arXiv e-prints : arXiv-2002.
[8] A. Aman, D.J. Reji, Environmental due diligence data: A novel corpus for training environmental domain NLP models, Data Brief 108579 (2022) http://dx.doi.org/10.1016/j.dib.2022.108579.