EnvBert: An NLP model for Environmental Due Diligence data classification

Afreen Aman a,*, Deepak John Reji b

a Ernst & Young, India
b Environment Resources Management, India

A R T I C L E I N F O

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

A B S T R A C T

EnvBert is a Natural Language Processing (NLP) package built on top of DistilBERT, focused on Environmental Due Diligence (EDD) 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 data used to train this model has been open-sourced and can be accessed at Mendeley. A custom embedding layer was added to improve the model's performance. The EDD model with custom vector representation constitutes EnvBert python package which can be installed using the package manager pip.

* Corresponding author.
E-mail addresses: afreen.aman@gds.ey.com (A. Aman), deepak.reji@erm.com (D.J. Reji).

EnvBert is a pipeline solution which has an Environmental Due Diligence classification model followed by custom word vectors which data as a part of environment site assessments in due diligence. EnvBert has the following capabilities:

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

EnvBert is a solution which has an Environmental Due Diligence classification model followed by custom word vectors which data as a part of environment site assessments in due diligence. EnvBert has the following capabilities:

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

# 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

2665-9638/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Please cite this article as: A. Aman and D.J. Reji, EnvBert: An NLP model for Environmental Due Diligence data classification, Software Impacts (2022) 100427, https://doi.org/10.1016/j.simpa.2022.100427.
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 inferred 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:

```bash
pip install EnvBert
```

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 Mendele [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