Data in Brief 45 (2022) 108579

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Environmental due diligence data: A novel corpus for training environmental domain NLP models

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

[a] Ernst & Young, Bangalore, Karnataka, India
[b] Environment Resources Management, Kannur, Kerala, India

ARTICLE INFO

Article history:
Received 1 August 2022
Revised 20 August 2022
Accepted 2 September 2022
Available online 7 September 2022

Dataset link: Environmental Due Diligence Data (Reference data)

Keywords:
Natural language processing
Environmental due diligence
DistilBERT
EnvBert
PyPI
Hugging face

ABSTRACT

This article takes a step in the direction of adapting existing Natural Language Processing (NLP) models to diverse and heterogeneous settings of Environmental Due Diligence (EDD). The approach we followed was to enrich the vocabulary of deep learning models with more data from environmental domain by collecting the data from open-source regulatory documents provided by Environmental Protection Agency (EPA) [1]. We used active learning and data augmentation methods to resolve the imbalanced classes and fine-tuned DistilBERT on EDD data to develop environmental due diligence model which is hosted as an inference Application Programming Interface (API) on Hugging Face Hub. This model was packaged to predict EDD classes, determine relevancy and ranking, and allows users to fine tune the model to more EDD classes. This package, EnvBert is hosted on Python Package Index (PyPI) repository [2]. We anticipate that the rich EDD dataset that we used to train the model and create a package would help the users contribute for a variety of NLP tasks on EDD textual data, especially for text classification purposes. We present the data in raw format;

* Corresponding author.
E-mail address: afreen.aman@gds.ey.com (A. Aman).

https://doi.org/10.1016/j.dib.2022.108579

it has been open sourced and publicly available at https://data.mendeley.com/datasets/tx6vmd4g9p/4.

Specifications Table

Subject	Environmental Data Science, Natural Language Processing
Specific subject area	Environmental Due Diligence is carried out by firms for evaluating environmental conditions systematically and preparing remedial plans in accordance with the guidelines by regulatory bodies to help in determining the impact on environment, and aid in estimating the financial implications involved in its remedial measures. Environmental data though textually heavy is mostly unstructured and not consistently available. Majority of the available data is the result of environmental monitoring, either in situ or via remote sensing [3,4]. Of late, like all domains environmental firms are also leveraging their data, processes, and applications to derive insights from unstructured text-based data.
Type of data	Table - Textual Data
How the data were acquired	The training data was handpicked from documents from EPA website [1].
Data format	Raw
Description of data collection	The training data consisted of 1576 sentences, spread across the 12 labels. Text data is prone to noise during collection as well as in test environments; even the same classification model may show different performance in different datasets [5]. We have handled the noise in our model by training the model to identify not relevant sentences. The data collection was uneven for a few categories, we used active learning and data augmentation to overcome the hindrance.
Data source location	Primary data source: Regulatory documents from Environmental Protection Agency (EPA) website: Search for Superfund Decision Documents | US EPA Table 1.0 in Appendix A contains the list of documents from EPA from which data was obtained. The secondary data corresponding to the source has been placed in Mendeley.
Data accessibility	Repository name: Mendeley Data identification number: 10.17632/tx6vmd4g9p.4 Direct URL to data: https://data.mendeley.com/datasets/tx6vmd4g9p/4
Related research article	Not Applicable

Value of the Data

- This dataset is a unique and one-of-a-kind corpus of environmental due diligence. There are currently no open-source datasets or models available for NLP in this domain.
- This data will help data scientists in building new models for a range of tasks in the EDD domain. It can be used for training NER models to recognize and extract entities, train QnA models, perform topic modelling etc.
- This dataset can serve as a reference/ training material for EDD analysts.
- Environmental Consultants/ Scientists/ Specialists/ Engineers/ Researchers, Site assessment and management Firms, Regulatory bodies etc. can access the data pertaining to existing categories and can contribute to the dataset by adding more categories.

1. Data Description

The dataset consists of data for 11 EDD categories and a ‘Not relevant’ category.

1. Remediation Standards: Numerical standards that define the concentrations of contaminants that may be permitted to remain in any environmental media.
2. Extent of Contamination: Spread of contamination and the level to which the media is contaminated.
3. Depth to water: The depth below the surface at which the water is found.
4. Groundwater-Surface water interaction: The interaction between the surface water and groundwater to detect the flow of contaminants.
5. GW Velocity: The velocity of the flow of water as well as contaminants. This category also includes details like hydraulic gradient and conductivity.
6. Geology: The physical features of the site and their underlying geology such as rocks, faults, minerals, and groundwater.
7. Contaminated media: Media such as soil, sediments, rocks, groundwater, surface water etc that has been affected by a release of contaminants.
8. Remediation Activities: Testing, design, treatment, removal, corrective action or other similar activities undertaken pursuant to Environmental Laws to address environmental contamination.
9. Remediation Goals: The goals set to ensure that the residual risks that remain at the site after clean-up will be within some specified limit of acceptability.
10. Source of contamination: Origin of a hazardous substance which is the first part of an exposure pathway.
11. Contaminants: Hazardous substances polluting the environment.
12. Not Relevant: The data that doesn't belong to the above EDD categories.

An example of the dataset is presented in Fig. 1.

content	Label
Elevated concentrations of chromium, copper, lead, nickel, silver, tin, and cyanide were found in the soil adjacent to the facility.	Contaminants
The remedial action selected is only a part of a total remedial action (interim remedy) and the final remedy will attain the ARAR upon its completion.	Remediation Activities
As with ground water, soil contamination should be documented in both vertical and horizontal directions.	Extent of contamination
Upon completion of ground water treatment, the water would be discharged offsite to the nearby tributary of North Creek.	Groundwater-Surfacewater interaction
Sampling depths were tilled depth for gardens (generally 0–12 inches), 0–3 inches for disturbed areas (e.g., animal activity areas), 0–1 inch for other residential soils, and 0–6 inches for beaches.	Depth to Water
Registry (ATSDR) must conduct a health assessment for every site proposed for inclusion on the NPL.	Remediation Goals
For example, preliminary treatment costs for contaminated soil can be calculated for various contaminant types and volumes.	Contaminated media
A potential source of MEC remains at the site as a result of historic military use.	Source of contamination
Constituents that exceed the MCL (or RSL/PRG in the absence of a MCL) thresholds are further evaluated in the refinement of COCs step (i.e., uncertainty discussion).	Remediation Standards
Site Mapping/Site Dynamics Map site and determine topography; determine site boundaries, drainage patterns, and other geophysical features.	Geology
A velocity of 2.1 x 10-1 cm/sec with a geometric mean of 7.4 x 10-2 cm/sec.	GW Velocity
Its teams have been unable to access the site due to restrictions in place to slow the spread of the coronavirus, he said.	Not Relevant

Fig. 1. Extract from the dataset.

2. Experimental Design, Materials and Methods

Quality data labelling is a key factor in a supervised model's performance. We prepared a curated training dataset by labelling sentences manually. The data was collected from regulatory documents and categorized into 12 classes. The data is retained in raw format i.e., cleaning, stemming, lemmatization or any type of pre-processing has not been applied after data collection. The articles contain some symbols, punctuations and digits. The distributions of the dataset per category (aka label) in terms of count are depicted in Fig. 2.

The data volume for certain categories was relatively low due to minority class instances. Training a classification model with limited data may cause class imbalance which affects the

label	count
Contaminants	158
Contaminated media	147
Extent of contamination	128
Geology	153
Groundwater-Surfacewater interaction	158
GW Velocity	124
Depth to Water	148
Remediation Activities	158
Remediation Standards	129
Remediation Goals	157
Source of contamination	122
Not Relevant	142

Fig. 2. Corpus statistics table.

predictions by favoring dominant categories [6]. Using oversampling to address this may result in overfitting of the model's predictions. This calls for a balanced approach. We used data augmentation technique, carefully reviewing the augmented data and adjusting it as needed. Contextual word embeddings with the bert-base-uncased model were used to generate/rephrase the existing data. The generated output retains the same context with a different sentence structure. Data Augmentation improves the diversity of the training dataset, thereby the model can better generalize to unseen testing data [7]. Other approaches to data augmentation include static word embeddings, back translation, text generation etc. Since contextual embeddings have achieved state-of-the-art against many language tasks [8], they were the best choice for this use case.

We also used active learning approach to further increase the size of the dataset for labels with limited data by predicting over new set of data. EnvBert was used to collect the environmental data with a high probability score for these labels. This learning process reduces the human annotation effort by only requiring manual review to ensure accuracy [9]. It was an iterative approach, retraining the model with high-quality data after each review. DistilBERT was finetuned on this meticulously curated dataset to create EDD model.

This EDD model with custom embeddings served as the pipeline components in the EnvBert Package. A flowchart detailing the data creation is presented in Fig. 3.

This EDD model is registered and deployed on Hugging Face Hub and it can be accessed at https://huggingface.co/d4data/environmental-due-diligence-model. EnvBert package is hosted on PyPI repository and can be accessed at https://pypi.org/project/EnvBert/ [2].

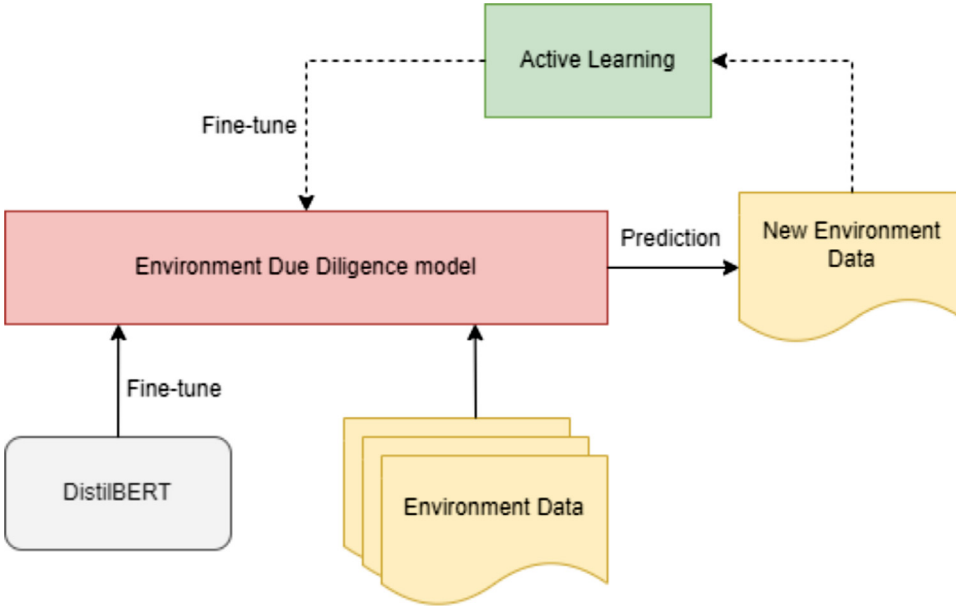

Fig. 3. Data process overview.

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.

Data Availability

Environmental Due Diligence Data (Reference data) (Mendeley Data).

CRediT Author Statement

Afreen Aman: Conceptualization, Methodology, Data curation, Writing – review & editing; **Deepak John Reji:** Conceptualization, Data curation, Software, Writing – review & editing.

Ethics

The data was collected from documents at EPA website. EPA permits the use of documents for non-commercial, scientific and educational purposes. EPA data disclaimers can be accessed through the link: https://www.epa.gov/web-policies-and-procedures/epa-disclaimers.

Table 1.0 in the Appendix section contains the name of the documents from which the data was used.

Appendix A: Tables

Table 1.0. Source documents from EPA

Sl. No.	Document Name
1	Guidance for Conducting Remedial Investigations and Feasibility Studies Under CERCLA, Interim Final, OSWER Directive 9355.3-01
2	SUPERFUND TASK FORCE QUARTERLY REPORT FY19 Q1
3	Smart Scoping of an EPA-Lead Remedial Investigation/Feasibility Study \(RI/FS\) Fact Sheet - EPA 542-F-19-006
4	Guidance for Sample Collection for In Vitro Bioaccessibility Assay for Arsenic and Lead in Soil and Applications of Relative Bioavailability Data in Human Health Risk Assessment
5	Guidance for Sample Collection for In Vitro Bioaccessibility Assay for Arsenic and Lead in Soil and Applications of Relative Bioavailability Data in Human Health Risk Assessment - Appendix A
6	Guidance for Sample Collection for In Vitro Bioaccessibility Assay for Arsenic and Lead in Soil and Applications of Relative Bioavailability Data in Human Health Risk Assessment - Attachment A: FAQs
7	Superfund Radiation Risk Assessment Calculator Training
8	Superfund Community Advisory Group Toolkit for the Community, EPA 540-B-22-001
9	Superfund Community Advisory Group Fact Sheet for the Community
10	FY 2021 Superfund Accomplishments Report
11	Final Record of Decision for Operable Unit 14 Ellsworth Air Force Base, South Dakota
12	FINAL ACTION MEMORANDUM - NON TIME CRITICAL REMOVAL ACTION \(NTCRA\) TO ADDRESS OFF-BASE PFAS-CONTAMINATED MUNICIPAL AND RESIDENTIAL DRINKING WATER SUPPLY WELLS, FIRE TRAINING AREA-1 \(ASHUMET VALLEY\)
13	FINAL EXPLANATION OF SIGNIFICANT DIFFERENCES \(ESD\) - OPERABLE UNIT \(OU\) 1
14	FINAL RECORD OF DECISION \(ROD\) FOR MOCK VILLAGE MUNITIONS RESPONSE SITE, EPA OPERABLE UNIT \(OU\) 29
15	RECORD OF DECISION FOR THE RIVER CITY METAL FINISHING SUPERFUND SITE
16	THIRD FIVE-YEAR REVIEW REPORT FOR PALMER BARGE LINE SUPERFUND SITE JEFFERSON COUNTY, TEXAS
17	RECORD OF DECISION, REMEDIAL ALTERNATIVE SELECTION FOR THE LOWER THREE RUNS INTEGRATOR OPERABLE UNIT \(U\), SAVANNAH RIVER SITE.
18	RECORD OF DECISION, US FINISHING/CONE MILLS, OPERABLE UNIT 1 SUPERFUND SITE, GREENVILLE, GREENVILLE COUNTY, SOUTH CAROLINA.
19	RECORD OF DECISION, US FINISHING/CONE MILLS, OPERABLE UNIT 1 SUPERFUND SITE, GREENVILLE, GREENVILLE COUNTY, SOUTH CAROLINA.
20	REPORT: MEETING COMMUNITY NEEDS, PROTECTING HUMAN HEALTH AND THE ENVIRONMENT: ACTIVE AND PASSIVE RECREATIONAL OPPORTUNITIES AT ABANDONED MINE LANDS \(AML\)
21	FACT SHEET: THE REMEDIAL INVESTIGATION: SITE CHARACTERIZATION AND TREATABILITY STUDIES, OSWER 9355.3-01FS2
22	FACT SHEET: THE FEASIBILITY STUDY: DEVELOPMENT AND SCREENING OF REMEDIAL ACTION ALTERNATIVES, OSWER 9355.3-01FS3
23	QUICK REFERENCE FACT SHEET: THE FEASIBILITY STUDY: DETAILED ANALYSIS OF REMEDIAL ACTION ALTERNATIVES, OSWER 9355.3-01FS4
24	QUICK REFERENCE FACT SHEET: TREATABILITY STUDIES UNDER CERCLA: AN OVERVIEW, OSWER 9380.3-02FS
25	MEMO REGARDING THE LAND DISPOSAL RESTRICTIONS AS RELEVANT AND APPROPRIATE REQUIREMENTS FOR CERCLA CONTAMINATED SOIL AND DEBRIS OSWER 9347.2-01
26	TECHNICAL BACKGROUND DOCUMENT - PART 2 - DEVELOPMENT OF PATHWAY-SPECIFIC SOIL SCREENING LEVELS
27	SOIL SCREENING GUIDANCE: USER'S GUIDE
28	CONDUCTING REMEDIAL INVESTIGATIONS/FEASIBILITY STUDIES FOR CERCLA MUNICIPAL LANDFILL SITES OSWER 9355.3-11 EPA 540-P-91-001
29	GUIDANCE FOR CONDUCTING TREATABILITY STUDIES UNDER CERCLA - FINAL OSWER 9380.3-10 EPA 540-R-92-071A
30	FEASIBILITY STUDY ANALYSIS FOR CERCLA MUNICIPAL LANDFILL SITES OSWER 9356.0-03 EPA 540-R-94-081 PB95-963301
31	FACT SHEET - WATER QUALITY CREDITS AT FORMER LAND MINES: IMPROVING AMERICA'S WATER RESOURCES, RECLAIMING LOST LANDSCAPES
32	UNITED STATES ENVIRONMENTAL PROTECTION AGENCY MODEL GOOD SAMARITAN SETTLEMENT AGREEMENT AND ORDER ON CONSENT FOR REMOVAL ACTIONS AT ORPHAN MINE SITES

(*continued on next page*)

Sl. No.	Document Name
33	FACT SHEET: REMEDIAL INVESTIGATION BEGINS
34	FACT SHEET: REMEDIAL INVESTIGATION \(RI\) BEGINS
35	Record of Decision for Construction Debris Landfill (CDL) Operable Unit 5 (OU5) National Aeronautics and Space Administration (NASA) Langley Research Center (LaRC)
36	OU 2 RECORD OF DECISION \(ROD\)
37	OU 2 RECORD OF DECISION \(ROD\)
38	Fifth Five-Year Review Report for Fike Chemical Site July 2017
39	Five Year Review Recommended Template
40	RECORD OF DECISION \(ROD\)
41	RECORD OF DECISION FOR OU2 FOR THE TUTU WELLFIELD SITE
42	THIRD FIVE-YEAR REVIEW REPORT FOR THE CHEMICAL LEAMAN TANK LINES SITE
43	RECORD OF DECISION FOR OU4 FOR THE WELSBACH & GENERAL GAS MANTLE \(CAMDEN RADIATION\) SITE
44	EXPLANATION OF SIGNIFICANT DIFFERENCES FOR OU2 AND OU4 FOR THE ROCKAWAY BOROUGH WELL FIELD SITE
45	MEMORANDUM REGARDING THE CONTAMINATED SEDIMENTS TECHNICAL ADVISORY GROUP \(CSTAG\) RECOMMENDATIONS ON THE ALLIED PAPER, INC./PORTAGE CREEK/KALAMAZOO RIVER SUPERFUND SITE
46	NPL SITE LISTING NARRATIVE
47	RECORD OF DECISION FOR OU4 FOR THE MCGUIRE AIR FORCE BASE #1 SITE
48	EXPLANATION OF SIGNIFICANT DIFFERENCES [ESD] \(SIGNED\) - SOUTH ANDOVER SITE
49	LEIDOS - FINAL RECORD OF DECISION, SVAD-222, EXPLOSIVE BUILDING DECONTAMINATION, SAVANNA ARMY DEPOT ACTIVITY, SAVANNA, ILLINOIS\n
50	EXPLANATION OF SIGNIFICANT DIFFERENCES \(ESD\) \(SIGNED\) LITTLE SCIOTO RIVER OU1
51	EPA - EXPLANATION OF SIGNIFICANT DIFFERENCES \(ESD\) \(SIGNED\)
52	EXPLANATION OF SIGNIFICANT DIFFERENCES \(ESD\) \(SIGNED\) MACGILLIS AND GIBBS/BELL LUMBER POLE SUPERFUND SITE
53	INTERIM RECORD OF DECISION AMENDMENT \(ROD\) - ADAM'S PLATING SUPERFUND SITE
54	RECORD OF DECISION \(ROD\) - MILFORD CONTAMINATED AQUIFER SUPERFUND SITE
55	RECORD OF DECISION FOR ALLIED PAPER/PORTAGE CREEK/KALAMAZOO RIVER, OPERABLE UNIT 5, AREA 3
56	FIFTH FIVE-YEAR REVIEW REPORT \(SIGNED\) - MOTOR WHEEL INC - 2022
57	EXPLANATION OF SIGNIFICANT DIFFERENCES \(ESD\) \(SIGNED\) FISHER-CALO SUPERFUND SITE
58	SIXTH FIVE-YEAR REVIEW REPORT \(SIGNED\) - E.H. SCHILLING LANDFILL - 2022
59	SIXTH FIVE YEAR REVIEW REPORT \(SIGNED\) - OAK GROVE SANITARY LANDFILL - 2022
60	SIXTH FIVE YEAR REVIEW REPORT \(SIGNED\) - 2022

References

[1] Environmental Protection Agency (EPA): https://www.epa.gov/.
[2] EnvBert: https://pypi.org/project/EnvBert/
[3] F. Kogan, A. Powell, O. Fedorovk, Use Of Satellite and *In-Situ* Data To Improve Sustainability eds., Springer, 2011.
[4] R.M. Darbra, N. Pittam, K.A. Royston, J.P. Darbra, H. Journee, Survey on environmental monitoring requirements of European ports, J. Environ. Manage. 90 (3) (2009) 1396–1403.
[5] J. Du, Q. Chen, Y. Peng, Y. Xiang, C. Tao, Z. Lu, ML-Net: multi-label classification of biomedical texts with deep neural networks, J. Am. Med. Inform. Assoc. 26 (11) (2019) 1279–1285.
[6] F. Thabtah, S. Hammoud, F. Kamalov, A. Gonsalves, Data imbalance in classification: experimental evaluation, Inf. Sci. 513 (2020) 429–441.
[7] B. Li, Y. Hou, W. Che, Data Augmentation Approaches in Natural Language Processing: A Survey, AI Open, 2022.
[8] M.E. Peters, M. Neumann, L. Zettlemoyer, W.T. Yih, Dissecting contextual word embeddings: architecture and representation, arXiv (2018) preprint arXiv:1808.08949.
[9] M. Prince, Does active learning work? A review of the research, J. Eng. Educ. 93 (3) (2004) 223–231.