Office of Environment, Health, Safety and Security
Operating Experience Summary

OES 2023-01-P3 February 2023

Trending Analysis and Machine Learning (ML) Part 3: Lessons Learned on ML Tools Design, Development and Use

Introduction

This Operating Experience Summary (OES) presents Part 3 of a four-part series and provides information about the lessons learned from the design, development and deployment of data analytics and machine learning (ML) tools used in support of the most recent U.S. Department of Energy (DOE) Office of Environment, Health, Safety and Security (EHSS) fire protection data trends report issued for calendar years (CY) 2015-2019 (report link). The lessons in this OES will aid other DOE stakeholders who are developing similar applications to analyze environment, safety, and health (ES&H) data and may provide better understanding of the capabilities and limitations of these tools.

Background

Since 2018, DOE's Office of ES&H Reporting and Analysis (EHSS-23) has been developing applications of ML to improve the efficiency of analyzing ES&H data. Throughout this process, these tools are being used to analyze report data, aid ES&H practitioners in identifying potential issues, and in decision-making. The tools have various modules that allow for data visualization, advance search techniques (e.g., using document similarity), and group ES&H reports (e.g., using ML clustering analysis[1] and algorithms). The tools are available to DOE federal and contractor employees and EHSS-23 can conduct ML tools demonstrations for both DOE and non-DOE stakeholders.

Discussion of Lessons Learned

During the design, development, and deployment of the tools, the team identified various lessons learned that could aid other stakeholders:

- Limiting the scope of the initial project helps with incremental development and highlights early successes.
- Verifying results from an ML-based analysis with results from a non-ML based analysis helps to validate the performance and accuracy of ML-based results and increase trust in such tools.

The quality of your source data is an important factor. Known and unknown issues may impact the performance of models. The following are lessons and general data observations:

- It is important to understand historical changes on a report's assigned categories and classes (e.g., type of fire) and how these affect training data.
- It is important to identify database limitations, errors and other known issues that can affect performance of artificial intelligence (AI) models.
- In some cases, ML relies on those data points with highest frequency within the data. However, ES&H data may contain rare events of significance (i.e., low frequency high

[1] Clustering analysis is a machine learning method that divides data points into batches or groups based on similarities or difference in their properties.

consequence). Consider integrating risk analysis of less-frequent events with high consequence potential.

The ES&H tools leverage natural language processing[2] (NLP) and open-source data science libraries. Depending on the method used, NLP can include various steps including but not limited to text normalization, removal of “stop” words, model development, and deployment. The use of particular “stop” words can introduce complications, so following lessons learned are specific to the use of “stop” word libraries.

- The words ’no’ and ’not’ are typically considered “stop” words in NLP libraries. The removal of these words from the analysis, however, may change the context of some of the results. For example, if a narrative has ‘no fire’ or ‘no injury’, removing the word ‘no’ will completely change its meaning.
- Some existing libraries may include “stop” words important to the data being analyzed. For example, one of the libraries included the word ‘fire’ as a stop word. Without correcting, this would cause problems when using this library to analyze fire protection data.
- Application of “stop” words needs to be consistent throughout text preprocessing, normalization, vectorization, and analysis functions.

When using ML clustering, use of various algorithms (e.g., k-means[3], DBSCAN[4]) can help minimize the limitations and leverage the strengths of individual algorithms. For example, in the fire protection trend analysis discussed in Parts 1 and 2 of this OES series, DBSCAN was used to identify clusters and terms including outlier reports. The k-means algorithm cannot be used to identify outliers, but it is able to identify actionable clusters and terms that were missed by the DBSCAN algorithm. Running both algorithms may just add a few minutes, but it can reduce the risk of missing important clusters and terms.

Summary

These lessons learned and considerations should help improve the outputs of data analytics and ML tools and models. Important points include:

- Develop approaches to verify the outputs of your ML models, including those events that have low frequency but may have a high impact and consequence.
- Understand the history and limitations of your data and how changes affect ML models.
- Exercise caution when using stop word libraries in analysis that use NLP.

The Office of Environment, Health, Safety and Security (EHSS), Office of ES&H Reporting and Analysis publishes the Operating Experience Summary to promote safety throughout the Department of Energy (DOE) Complex by encouraging the exchange of lessons-learned information among DOE facilities.

Please contact Felix Gonzalez at 301-903-9311 or Felix.Gonzalez@hq.doe.gov for questions regarding this OE Summary.

OES Series: Trending Analysis and ML

This OES is part of a series of articles focused on the application of trending analysis and machine learning to DOE fire protection data and other ES&H trending analysis reports. The series includes the following four parts:

Part 1: *DOE Fire Protection Trends*
Part 2: *DOE Fire Protection ML Text Clustering*
Part 3: *Lessons Learned on ML Tools Design, Development and Use* **(this OES document)**
Part 4: *References on Artificial Intelligence Best Practices and Principles*

[2] Natural language processing is a branch of artificial intelligence where computers are enabled to understand text and spoken words.

[3] k-means partitions unlabeled data into a certain number (i.e., “k”) of distinct groupings.

[4] Density-Based Spatial Clustering of Applications with Noise (DBSCAN)