Data Science Applications in Safety Science

January 28, 2022

E. Andrew Kapp, Phd, CSP, CHMM

UNDERWRITERS LABORATORIES®

Data Science Applications in Safety Science

Introduction

Safety can be defined as the freedom from danger, risk, or injury. Safety science, then, is the systematic study of the structure and behavior of the physical world to understand and mitigate danger, risk or injury through observation and experimentation. Lapses in safety can result in property loss, unintentional injuries, and death; most safety issues are preventable. Unintentional injuries are the cause of more than 3 million deaths globally (Global Burden of Disease Collaborative Network, 2019), a true burden on society.

Data is plentiful in the field of injury prevention and safety science. Governments, universities, and safety organizations around the world collect millions of records related to safety each year. Hospitalization records, recalls, workplace incidents, reports of unsafe products and regulatory enforcement actions provide a vast record of actual and potential safety incidents. The improvement in the collection and analysis of these safety incidents- particularly using the structured data in the records- has fueled improvement in safety for decades. From structured data, demographic information, product information and injury/damage information can be readily extracted and analyzed. This provides trends, incidence rates and many additional useful insights for safety researchers.

Typically, these safety incident records include a field for additional detail not captured in the pre-defined, structured data. These fields take the form of a narrative description of the incident, and, increasingly, pictures or videos related to the incident. These unstructured data are largely an untapped resource due to the resource intensive efforts required to review the data. Most research using the unstructured data has required individual researchers to read, parse and code the information into additional structured data for subsequent analysis. These efforts become impractical when seeking to analyze large sets of incident records.

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and computer systems to extract knowledge and insights from structured and unstructured data. The growth of computing capabilities and their lower cost have fueled the expansion of data science into new applications, including safety science. Several noteworthy advances in data science – machine learning (ML) and natural language processing (NLP) – are particularly useful in developing new safety insights from the large volumes of safety incident data, including incident narratives.

This paper serves as a review of several projects undertaken by the Underwriters Laboratories Data Science team that apply these novel data science approaches to traditional safety issues. This paper provides an overview of NLP and ML and examples of the application of these techniques. Project summaries are provided of the UL Data Lake and two of its ML features, the recommendation engine and safety marks detection algorithm; a second application of the recommendation engine, the Horizontal Requirements Analysis; and several classification projects using supervised ML and transfer learning – Furniture Tipover, ICS Code Classification and Hazard Classification. By sharing these projects, including methods and algorithms, additional opportunities to advance safety science through data science may come to light.

An overview of natural language processing and machine learning

Natural Language Processing (NLP) is the conversion of text and voice data into a format that allows computers to perform additional manipulation, computation, and analysis. Machine Learning (ML) refers to a subfield of artificial intelligence (AI) where the objective is to fit a set of data to a model to uncover patterns. ML can work with text, numeric or even image data. In the case of text data, the raw text must be processed and converted into numerical data in order to be analyzed via NLP. The Data Science team traditionally has used one of two methods for NLP: Term Frequency and Inverse Document Frequency (TF-IDF) and Bidirectional Encoder Representations from Transformers (BERT) and its derivatives.

Natural Language Processing

Pre-processing

Before many NLP algorithms can be employed, the text must be pre-processed to transform the useful information into a format which the algorithm can use. The first step in the preprocessing of text is consolidation. Here the useful text fields from varying locations are transferred into individual entries in a single text document. Tokenization follows, where the stream of text is broken down into individual terms, words, symbols, or other meaningful elements called tokens. Next is part of speech (POS) tagging where individual words are labeled according to the part of speech such as noun, verb, adjective, adverb, etc. “Stop” words such as “I”, “a” or “the” that do not carry any meaning within the scope of the investigation are removed to speed the preprocessing and streamline the subsequent analysis. Lemmatization follows, converting words back into their common root form (or lemma). For example, the lemma of the words “is”, “was”, “am”, and “being” are all converted to “be”. Following these steps, we have clean and concise text which can be transformed into numeric data that ML can process through TF-IDF or other algorithms.

Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF is a technique to quantify words in text by computing a weight to each word which signifies the importance of the word in the document and corpus. A corpus is a large collection of text data that is used to train machine learning systems. Term Frequency (TF) indicates how often a term occurs in a single record. Inverse Document Frequency (IDF) specifies how common or rare a term is in a given corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general use. TF-IDF values are then normalized so that the length of the text entries (now transformed into numeric data) does not have any biasing effect on the performance of the NLP algorithms. We are interested in the meaning of the text and finding patterns among these meanings, and the particular length of any text doesn't contribute to this.

Bidirectional Encoder Representations from Transformers (BERT)

BERT, developed by Google, is an open-source software that provides a pre-trained model for NLP. BERT was trained to understand language using content from Wikipedia and Books Corpus, which together contain over 3.2 billion words. Because BERT comes pre-trained, only tokenization, where text is broken down into individual terms or words, is required for pre-processing text. Unlike other encoders, BERT uses context aware encoding, meaning distinct vector representations are created for homonyms (words that are spelled and sound the same but have different meanings) and homographs (words that are spelled the same but sound different and have different meanings). For example, other encoders would have the same vector representations for the term “running” when used in the phrases “running a company” and “running a marathon”. Not so with BERT. BERT also utilizes bidirectional encoding to account for differing structures of sentences and negative phrasing. For example, the same vector representations are used for the term “guard” in “he did not use the guard” and “the guard was not used”.

Further advances on BERT have been developed and are being employed by the Data Science team, including RoBERTa (Robustly Optimized BERT), and DistilBERT. RoBERTa, developed by researchers at Facebook, improves the BERT base model by using dynamic masking, full sentence without next sentence prediction (NSP) loss, large mini-batches, and larger Byte-Pair Encoding (BPE) techniques. These optimization methods are beyond the scope of this paper. Importantly, the combination of these implementation techniques, the RoBERTa language model achieved significant performance improvement from the BERT base model in most NLP downstream tasks by 2% to 4% (Liu, et. al., 2019). DistilBERT, another variation of BERT, reduces computational expense (time and computing power) in training the language representation model by using a smaller pre-trained BERT model for general purposes, while retaining BERT’s performance and its capability of fine-tuning for a specific task. DistilBERT was found to be 60% faster with 40% fewer parameters than the original BERT and RoBERTa models while preserving 97% of language understanding capabilities (Sanh, et. al., 2019).

Machine Learning (ML)

Following natural language processing via TF-IDF or BERT (or one of the BERT derivatives), analysis of the numeric vector data can begin with either supervised or unsupervised ML algorithms. Supervised ML utilizes an algorithm that ‘teaches’ a function to map an input to an output based on example input-output pairs. A training data set is first labeled to provide examples of the correct classification. This data set is used by the algorithm to determine the function and gauge correct usage. A validation data set is used to test the performance of the algorithm. A validation data set is also labeled with the correct classification, but the “trained” algorithm is blinded to the labeling, analyzes the validation data and produces an independent prediction of the label. The results of the algorithm are compared against the pre-labeled data to determine the algorithms performance. A successful algorithm will then be able to correctly determine the correct output label a high percentage of the time.

Unsupervised ML involves the algorithm learning the function or patterns from unlabeled data without training or validation. During the learning phase, an unsupervised network tries to mimic the data it's given and uses the error in its mimicked output to correct itself (e.g., its weights & biases). This resembles the mimicry behavior of children as they learn a language. The interpretation of the patterns (outputs) from the algorithm is then left to the people.

Together, NLP and ML provide powerful tools for extracting and analyzing data from collections of text and images. By interpreting text or image data and identifying patterns in the data, our ability to make sense of otherwise overwhelming quantities of seemingly random data becomes possible. The remainder of this paper provides examples of our employment of these sense making techniques to enhance to the work of Underwriters Laboratories and the field of safety science.

Unsupervised Machine Learning Applications

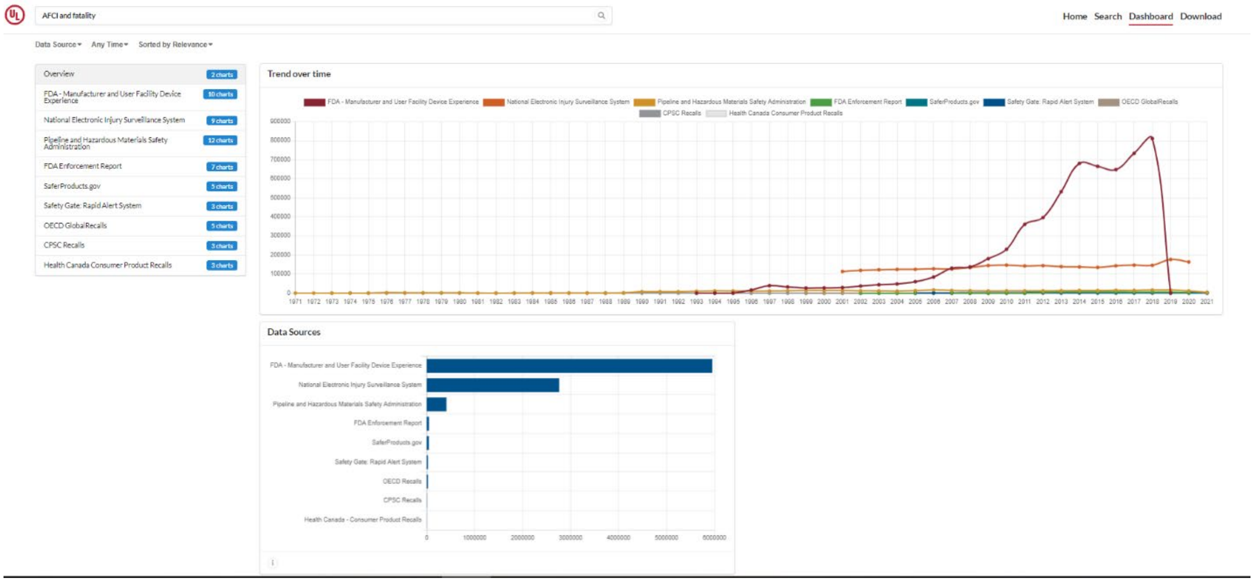
Safety Data Lake with Recommendation Engine

Background

The Safety Data Lake is a user-friendly research tool that can perform multiple tasks including data collection, data engineering, cloud computing, end-to end machine learning and data visualization. This platform is a one-stop service that enables users to:

- Search data across a disparate collection of publicly available data sets totaling more than 17 million records
- Identify and collect related incident records
- Perform basic analyses
- Review data visualizations of the results via a dashboard feature (Figure 1).

The recommendation engine feature of the Safety Data Lake makes ML driven algorithms available to all users to identify similar incidents/records to the record being viewed. This is accomplished through analysis of the unstructured data within the incident description, avoiding the time-consuming task of the user manually reading through many records to find incidents of interest.



AFCI and fatality

Home
Search
Dashboard
Download
Data Source ▾
Any Time ▾
Sorted by Relevance ▾
Overview 2 charts
FDA - Manufacturer and User Facility Device Experience 10 charts
National Electronic Injury Surveillance System 9 charts
Pipeline and Hazardous Materials Safety Administration 12 charts
FDA Enforcement Report 7 charts
SaferProducts.gov 5 charts
Safety Gate: Rapid Alert System 3 charts
OECD GlobalRecalls 5 charts
CPSC Recalls 3 charts
Health Canada Consumer Product Recalls 3 charts
Trend over time
FDA - Manufacturer and User Facility Device Experience
National Electronic Injury Surveillance System
Pipeline and Hazardous Materials Safety Administration
FDA Enforcement Report
SaferProducts.gov
Safety Gate: Rapid Alert System
OECD GlobalRecalls
CPSC Recalls
Health Canada Consumer Product Recalls
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
Data Sources
FDA - Manufacturer and User Facility Device Experience
National Electronic Injury Surveillance System
Pipeline and Hazardous Materials Safety Administration
FDA Enforcement Report
SaferProducts.gov
Safety Gate: Rapid Alert System
OECD Recalls
CPSC Recalls
Health Canada - Consumer Product Recalls
0
1000000
2000000
3000000
4000000
5000000
6000000

Figure 1: Sample Data Lake Dashboard Visualization

Purpose

The recommendation engine greatly improves the user experience with the Safety Data Lake by identifying the most similar incident records to the users' search parameters across the nine available databases of the Safety Data Lake. Without the recommendation engine, users would have to read through the individual incidents, screen the information provided, and use their judgement to identify similar records, an unreasonably time consuming and potentially error inducing process. The recommendation engine performs these tasks effortlessly, producing a set of suggested cases to the user that are the most relevant to their search specifications through the application of natural language processing and machine learning algorithms to the text data.

Method

Data Sources

More than 17 million public records spanning 1971 to the present were curated from government agencies dealing with public safety including the U.S. Consumer Product Safety Commission (CPSC), U.S. Department of Transportation (DOT) Pipeline and Hazardous Materials Safety Administration (PHMSA), U.S. Food and Drug Administration (FDA), the Organization for Economic Co-operation and Development (OECD), Health Canada and the European Commission (EU).

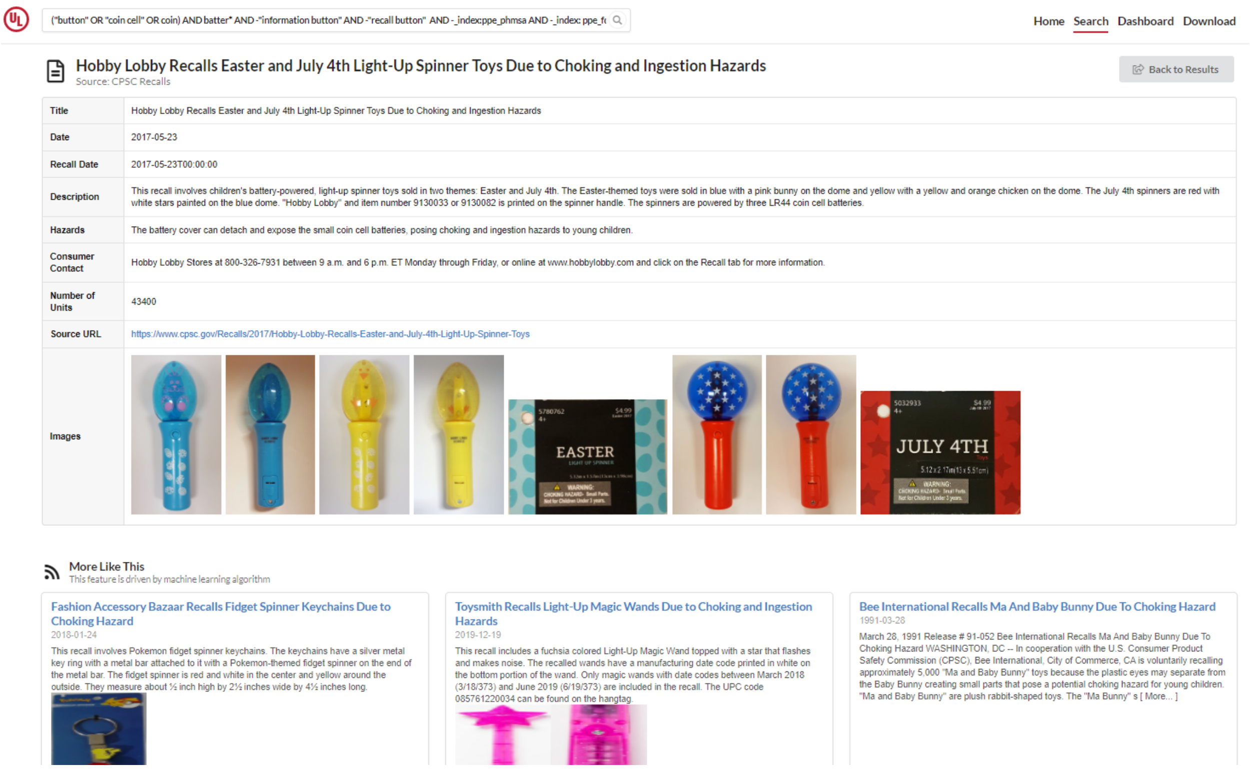
Specific data sets curated include:

- CPSC National Electronic Injury Surveillance System (NEISS)
- CPSC Recalls
- CPSC SaferProducts.gov (a public portal for reporting unsafe consumer products)
- EU Safety Gate: Rapid Alert System (formerly known as RAPEX)
- PHMSA Hazardous Materials Incident Reports
- FDA - Manufacturer and User Facility Device Experience (MAUDE)
- OECD Global Recalls
- Health Canada Consumer Product Recalls
- FDA Enforcement Report

All incident records where then pre-processed using the methods described in the overview section of this paper.

Machine Learning Implementation

A content-based recommendation algorithm was implemented for this use case. The algorithm is driven by K Nearest Neighbor (KNN), one of the unsupervised machine learning algorithms, to compare characteristics of an incident that the user is viewing at that time with entire database. Cosine similarity is assessed to determine which incidents have the same characteristics to the incident in which the user is interested. Once cosine similarity is calculated for each record, KNN sorts the similarity values in descending order and provide sorted incidents to users (Figure 2). To streamline the process for analyzing millions individual records, from preprocessing through implementation, we developed a pipeline using Apache Spark. Apache Spark is an analytics engine for large-scale data processing and ML development. Spark can process text on the fly and calculate cosine similarity against pre-processed text representation stored on the machine.



("button" OR "coin cell" OR coin) AND batter* AND -"information button" AND -"recall button" AND -_index:ppe_phmsa AND -_index: ppe_f
Home
Search
Dashboard
Download
Back to Results

Hobby Lobby Recalls Easter and July 4th Light-Up Spinner Toys Due to Choking and Ingestion Hazards

Source: CPSC Recalls

Title	Hobby Lobby Recalls Easter and July 4th Light-Up Spinner Toys Due to Choking and Ingestion Hazards
Date	2017-05-23
Recall Date	2017-05-23T00:00:00
Description	This recall involves children's battery-powered, light-up spinner toys sold in two themes: Easter and July 4th. The Easter-themed toys were sold in blue with a pink bunny on the dome and yellow with a yellow and orange chicken on the dome. The July 4th spinners are red with white stars painted on the blue dome. "Hobby Lobby" and item number 9130033 or 9130082 is printed on the spinner handle. The spinners are powered by three LR44 coin cell batteries.
Hazards	The battery cover can detach and expose the small coin cell batteries, posing choking and ingestion hazards to young children.
Consumer Contact	Hobby Lobby Stores at 800-326-7931 between 9 a.m. and 6 p.m. ET Monday through Friday, or online at www.hobbylobby.com and click on the Recall tab for more information.
Number of Units	43400
Source URL	https://www.cpsc.gov/Recalls/2017/Hobby-Lobby-Recalls-Easter-and-July-4th-Light-Up-Spinner-Toys
Images	

EASTER
LIGHT UP SPINNER
WARNING:
CHOKING HAZARD- Small Parts.
Not for Children Under 3 years.
JULY 4TH
WARNING:
CHOKING HAZARD- Small Parts.
Not for Children Under 3 years.

More Like This

This feature is driven by machine learning algorithm

Fashion Accessory Bazaar Recalls Fidget Spinner Keychains Due to Choking Hazard

2018-01-24

This recall involves Pokemon fidget spinner keychains. The keychains have a silver metal key ring with a metal bar attached to it with a Pokemon-themed fidget spinner on the end of the metal bar. The fidget spinner is red and white in the center and yellow around the outside. They measure about ½ inch high by 2½ inches wide by 4½ inches long.

Toysmith Recalls Light-Up Magic Wands Due to Choking and Ingestion Hazards

2019-12-19

This recall includes a fuchsia colored Light-Up Magic Wand topped with a star that flashes and makes noise. The recalled wands have a manufacturing date code printed in white on the bottom portion of the wand. Only magic wands with date codes between March 2018 (3/18/373) and June 2019 (6/19/373) are included in the recall. The UPC code 085761220034 can be found on the hangtag.

Bee International Recalls Ma And Baby Bunny Due To Choking Hazard

1991-03-28

March 28, 1991 Release # 91-052 Bee International Recalls Ma And Baby Bunny Due To Choking Hazard WASHINGTON, DC -- In cooperation with the U.S. Consumer Product Safety Commission (CPSC), Bee International, City of Commerce, CA is voluntarily recalling approximately 5,000 "Ma and Baby Bunny" toys because the plastic eyes may separate from the Baby Bunny creating small parts that pose a potential choking hazard for young children. "Ma and Baby Bunny" are plush rabbit-shaped toys. The "Ma Bunny" s [More...]

Figure 2: Data Lake with ML results for similar records.

Results & Implications

The Safety Data Lake is a full-stack data science tool that encompasses multiple tasks including data collection, data engineering, cloud computing development, end-to end ML development and data visualization. The addition of the recommendation engine offers the non-data scientist the ability to effortlessly apply NLP and ML algorithms to the incident descriptions of 17 million records to automatically search the disparate collection of unstructured text narratives to identify and collect similar cases. This ML driven feature helps users swiftly and easily discover relevant incident records without manually reading through nine entire databases, saving time and avoiding frustration. There have been multiple advanced ML algorithms released in recent years, and with these recent advances in ML, further exploration for our next improvement to enhance the language understanding capabilities lies ahead.

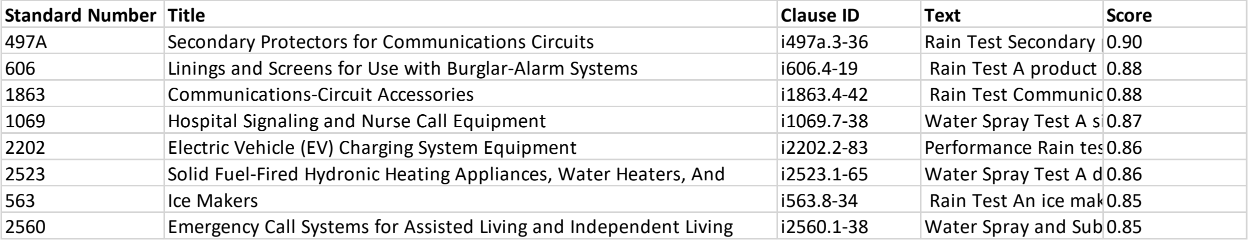
Horizontal Requirements Analysis

Background

Currently, a wide variation in test methods can be found within functionally equivalent tests required across UL standards. The Horizontal Requirements Analysis identifies these variations, informing opportunities to harmonize test methods across standards, or in circumstances where variations are justified, to document the rationale for such variations so as to ensure that the relevant hazards are addressed by way of the test method.

Purpose

The Horizontal Requirements Analysis project locates specific variations in the test requirements of functionally equivalent tests across more than 1700 UL standards (Figure 3). The details on the nature of the variations are then presented to subject matter experts (SME) who can subsequently determine if targeted changes to specific standard testing requirements are needed to achieve harmonization across standards, or if valid justifications for the variation exists that needs to be carefully documented.



Standard Number	Title	Clause ID	Text	Score
497A	Secondary Protectors for Communications Circuits	i497a.3-36	Rain Test Secondary	0.90
606	Linings and Screens for Use with Burglar-Alarm Systems	i606.4-19	Rain Test A product	0.88
1863	Communications-Circuit Accessories	i1863.4-42	Rain Test Communic	0.88
1069	Hospital Signaling and Nurse Call Equipment	i1069.7-38	Water Spray Test A s	0.87
2202	Electric Vehicle (EV) Charging System Equipment	i2202.2-83	Performance Rain tes	0.86
2523	Solid Fuel-Fired Hydronic Heating Appliances, Water Heaters, And	i2523.1-65	Water Spray Test A d	0.86
563	Ice Makers	i563.8-34	Rain Test An ice mak	0.85
2560	Emergency Call Systems for Assisted Living and Independent Living	i2560.1-38	Water Spray and Sub	0.85

Figure 3: Output of Horizontal Requirements Analysis

Method

Using UL Standards markup (XML and SGML) files (computer readable files with predetermined formats), we parsed the content to extract the standards' clauses. Standard SMEs determined the model text (test requirement) to provide a 'target' for the algorithm to determine clauses that are contextually matched with the model text. The sentence-transformer BERT was used compute dense vector representations for standards' clauses and the model text. Cosine similarity was computed to measure the distance between vector representations of the model text and each standards clause, with a cosine similarity score of 1 indicating that two vectors (i.e. clauses) are identical. The higher the cosine similarity score, the closer of vectors contextually are, and thus higher similarity between the model clause and the scored content. SMEs identified the threshold of cosine similarity to be more than 0.75 as relevant standard requirement with the model text.

Results & Implications

Applying ML in this project helps SMEs and Technical Committees to identify variations in test requirements enabling further harmonization of testing methods across standards. Additionally, by reducing variation in requirements, this project will result in better assurance that the standard mitigates the identified hazards resulting from the product or system. Further application of these methods could facilitate the development of an electronic database of requirements, test methods and rationale available to assist Standards Technical Panels (STPs) and Technical Committees in making timely and informed decisions regarding the appropriate test method to mitigate known hazards. The implications of this project extend to conformity assessment organizations, assisting them reducing the inefficiencies that come with maintaining the capabilities of running multiple variations of functionally equivalent tests.

Supervised Machine Learning Applications

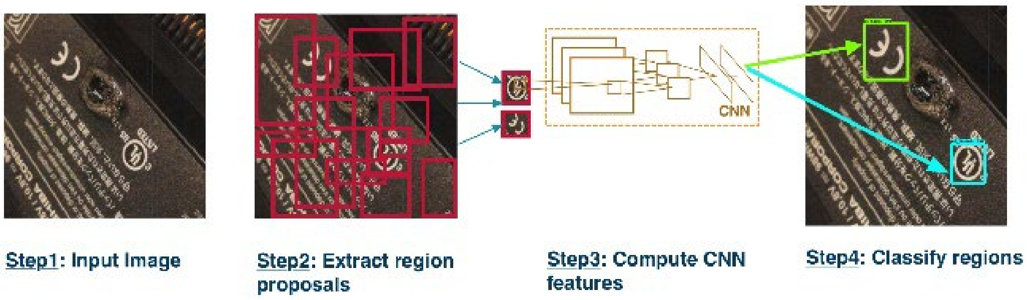
Safety Marks Recognition

Background

Manual review of images from product failure incident reports for conformity assessment organization marks is time consuming and can potentially lead to errors. The Safety Marks Recognition application is an ML function that automates this process. It is easy to use and access via the UL Safety Data Lake platform as one of the available filters. It allows users to easily filter incidents involving listed products through automatic screening and detection of safety marks in photographs accompanying incident reports in *SaferProducts.gov.*

Purpose

As stated above, the ML-driven object detection offered by the Safety Marks Recognition tool helps to reduce the time and labor previously required to identify the presence of safety marks on the images from consumer product failure reports gathered by the U.S. Consumer Product Safety Commission (CPSC). The pre-trained, region-based convolutional neural network algorithm, called Faster R-CNN, is capable of recognizing common objects such as cats, dogs, humans, and cars, for example. Through additional, specific training of the model, it is able to detect safety listing marks with more than 85% accuracy (Figure 4).



Step1: Input Image

Step2: Extract region proposals

Step3: Compute CNN features

Step4: Classify regions

Figure 4: UL Safety Marks Recognition Model Process

Method

To build the training and test sets, curated images from *SaferProducts.gov* were manually screened to identify safety marks by the UL Market Surveillance Team. Safety marks were manually labeled in each image. The labeled images of products bearing the safety marks were then split into training and test sets, with about 500 labeled images in the training set and about 150 images in the test set.

The first step in training is to teach the model where to look on the image to locate a possible mark by storing the coordinates (x,y) of the boundary boxes in the images with class/type of the box. The UL Safety Marks recognition model was then trained to specifically recognize safety marks using the labeled training data set. The model was validated by blindly applying the algorithm to the labeled images of the test set. To expedite the training, Faster R-CNN was used as a pre-trained model to identify basic objects as a baseline to train our model to specifically identify UL mark. After baselining the algorithm using the UL Mark, the safety mark detection algorithm was expanded to recognize a greater variety of safety marks (e.g., CSA, ETL and CE), as well as applying the algorithm to other databases that have product images.

Results & Implications

The safety mark detection was deployed in the Data Lake. It helps users find consumer incidents that involve certified products with greater ease and efficiency. The model was able to detect UL listed marks with more than 85% accuracy.

Furniture Tip-Over Classification

Background

The CPSC estimates that on an average of 25,000 emergency room visits and 571 fatalities occur each year from furniture tip-over incidents. The ability to distinguish furniture tip-over events from other furniture related incidents is vitally important when conducting research with incident reports to determine corrective measures. A new method using the combination of NLP techniques and supervised ML algorithms to analyze the incident narratives was developed to streamline incident classification of true tip-over incidents, greatly reducing time and effort.

Purpose

Through the use of NLP techniques and supervised ML algorithms, the identification of true furniture tip-over incidents from incident narratives can be automated reducing the time and labor of reading individual reports while achieving a high degree of precision.

Method

Beginning in 2017, incident narratives collected from the CPSC were classified manually to identify furniture tip-over events, but the labor-intensive process proved infeasible for continued use. TF-IDF, one of NLP techniques, was initially applied and then a Naïve Bayes algorithm explored to classify narratives. None of these approaches yielded satisfactory results. Newer language models and ML algorithms have been recently released that appeared promising for achieving better results.

Accordingly, we designed an experiment to evaluate several model variants to determine the most useful combination of techniques for the task of classifying furniture injury incident narratives into tip-over and non-tip-over categories. The data from the 2017 manual classification project was used for the experiment and it contained 2 classes of events: tip-over and non tip-over. The tip-over class contained 1123 incidents and the non tip-over class contained 7470 incidents. We split the dataset into random training and test sets. The training and test sets contained 6874 (80%) and 1719 (20%) records respectively.

Our base model architecture was derived from a pretrained RoBERTa model. We modified the RoBERTa classification by adding layer normalization (LayerNorm) which enables smoother gradients, faster training, and better generalization accuracy (Xu, Sun, Zhang, Zhao, & Lin, 2019), followed by a linear layer for classification on top of the base RoBERTa+LayerNorm model (Figure 5).

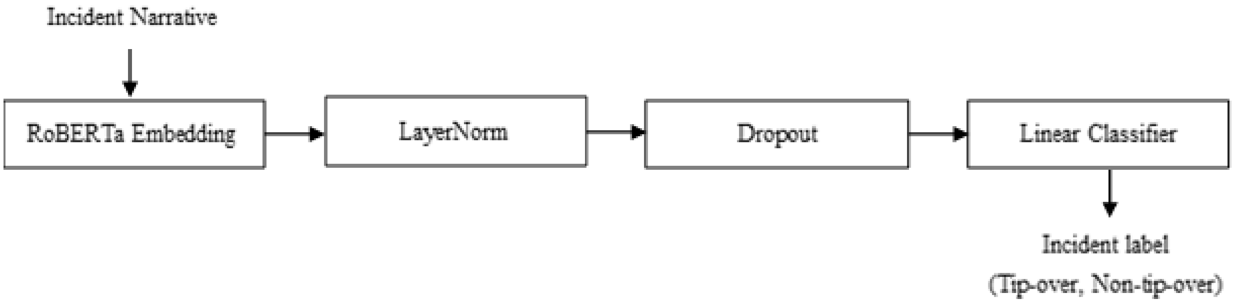


Figure 5: Architecture of RoBERTa+LayerNorm Classification Model

We adopted a transfer of learning technique to leverage pretrained models for the specific domain of injury narratives from NEISS. The pretrained language models, Bidirectional Encoder Representations from Transformers (BERT), optimized BERT (RoBERTa) and light-weight BERT (DistilBERT) were then used to preprocess free text and to fine tune the pretrained models for text classification task. The models using the default classification models from Transformers (Wolf et al., 2019) were run for comparison to RoBERTa+LayerNorm to find the best model for our use. For fine-tuning the process, all the models were trained on the same parameters for performance comparison. They were trained for 5 epochs with the batch size of 24. Adam from Kingma and Ba (2014) is used for optimization of the model with the learning rate of 1e-05. We used Transformers from HuggingFace (Wolf, et al., 2019) as a framework and Python language for implementation ML models.

Each model was evaluated for the furniture tip-over classification task using precision, recall and F1 score generated by a Python package called Scikit-learn. We also assessed each model on macro and class-wise levels to see how well they predict individual classes.

Results & Implications

Successfully using NLP and ML in classification helps improve workflow, increasing efficacy and reduce human efforts. Using incident narratives from the NEISS database from 2010 to 2015 we tested ML algorithms' abilities to classify whether an incident was a furniture tip-over or not. The performance of the models is reported in Table 1. RoBERTa with LayerNorm outperformed the original DistilBERT, BERT and RoBERTa classification models from HuggingFace. RoBERTa + LayerNorm was able to achieve a precision of 97%, a recall of 98% and an F1 score of 97%. Precision is the number of true positives divided by the number of true positives and false positives. Recall is the number of true divided by the number of true positives and the number of false negatives. The F1 score gives the balance between the precision and the recall: 2*((precision*recall)/(precision+recall)). This indicates an improved accuracy of the RoBERTa + LayerNorm over the BERT , DistilBERTand RoBERTa classification models of 3 to 5 %.

Model	Precision	Recall	F1
DistilBERT	0.92	0.92	0.92
BERT	0.92	0.95	0.93
RoBERTa	0.94	0.94	0.94
RoBERTa + NormLayer	0.97	0.98	0.97

Table 1: Classification Report - Macro Averaging Scores

Standards Classification: Establishing ICS Code & Covered Hazards Classification

Background

This exploratory project was designed to apply ML algorithms to label UL standards according to two established classification schemes to improve the ability to search the UL standards portfolio. By classifying standards according to these established systems, our standards portfolio can be more easily searched and relevant standards of interest related to the established classification criteria identified.

Purpose

This initiative was to classify UL standards under two established conventions: The International Classification for Standards (ICS) and CPSC National Injury Information Clearinghouse hazard classification (Hazard Type). ICS is a hierarchical classification covering 40 fields of activity in standardization managed by the International Organization for Standardization (ISO) and used in catalogues of international, regional and national standards as well as other normative documents. Hazard Type is a labeling system combining the object or substance, and the event that produced an adverse outcome from a consumer product. The addition of these two coding systems to the standards in our portfolio will allow stakeholders to quickly sort and select UL standards related to their specific interests. This tool will be of particular benefit to stakeholders searching for specific standards related to a prescribed topic.

Method

ICS Code classification

IEC and ISO portfolios, including over 40,000 abstracts, titles and ICS codes were used as input for ICS code classification algorithm. 35 of the 40 Level 1 ICS Codes were represented in the data. 3,626 National Standards of Canada scraped from SCC catalog include abstracts, titles with ICS codes were used as validation data.

The transfer learning technique was also adopted to leverage a pretrained model (BERT) to the specific domains of ICS and Hazard Type codes.

- ISO and IEC data were split into training and testing sets with the portion of 80% and 20% of total ISO and IEC data respectively.
- Trained multi-classification model on abstracts and titles using ICS level 1 code as a target label.
- Validated model on Canadian standards.
- Applied best performing model to UL Standard scopes and titles.

Covered Hazards Classification

CPSC Clearinghouse data (2011 to 2019) that includes 220,441 records with incident description and hazard were used for hazard types classification. There were 35 hazards in the data.

- Consumer products data were split into training (80%) and testing (20%) sets.
- Trained multi-classification model created with pre-trained DistilBERT on incident descriptions using hazard as a target label.
- Applied best performing model to UL Standard scopes and titles.

Results & Implications

The optimized BERT (RoBERTa) provided 84.71% accuracy score with 84.13% precision score and 84.71% of recall score for 35-class classification of ICS Level 1 codes. The lightweight BERT (DistillBERT) offered 85.49% accuracy score with 83.98% precision score and 85.49% of recall score for 35-class hazard classification of incidents. In addition to the validation of the model by test data, a sample of the output data was reviewed by Standards SMEs for accuracy. SMEs agreed with the Level 1 ICS code in 77% of the predictions and the Primary Hazard in 83% of the predictions. These results are in line with expectations for the prediction given the variables involved – subjectivity of classification, multiple and overlapping classifications, and subjectivity of SMEs. This project will enable users to find standards within categories in which they are interested, as well as to assign ICS codes and hazard classifications when new standards are developed.

Conclusion

The continued explosion of data across all domains of science presents a unique opportunity for the next several years. These large volumes of data contain information and insights that will allow safety professionals to develop intervention strategies to reduce unintentional injuries. However, the volume of data presents significant challenges to researchers and practitioners who must gather, process, analyze and act on the data to uncover the insights. Data science algorithms and methods, such as natural language processing, image detection, machine learning, and classification algorithms, provide the necessary tools to overcome the data volume challenge.

NLP approaches using TF-IDF and BERT allow researchers to mine the unstructured incident narrative text to determine key characteristics of incidents without a human reading the narratives. Classification algorithms group incident records based on these characteristics (e.g., furniture tip-over) with high accuracy and precision. This will allow the researcher to focus on the targeted mitigation of furniture instability rather than the classification problem.

Standards are written with the intent of mitigating risks to safety, security and sustainability. Standards Development Organizations (SDO) around the world have large portfolios of standards that contain a very large corpus of content that have been written over decades. Classifying these documents based on their content is a time and resource intensive endeavor. Using NLP and ML, the Data Science Team has developed methods to transfer the learnings from incident databases and previously classified standards to automatically process the content of standards for classification. These methods, after further refinement, may enable faster, more consistent and more efficient classification of standards based on existing and new classification models.

Finally, the Data Science Team has proven that algorithms and methods developed for applications that are vastly different from safety science can be adapted to successfully solve safety challenges. These algorithms and methods are being adapted, scaled, and productized to make safety research efforts more effective and efficient. Ultimately, data science is advancing Underwriters Laboratories' mission of working for a safer world.

References

Global Burden of Disease Collaborative Network. *Global Burden of Disease Study 2019 (GBD 2019) Results*. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2020. Available from http://ghdx.healthdata.org/gbd-results-tool.

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach*. arXiv preprint arXiv:1907.11692.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019*). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). *Huggingface's transformers: State-of-the-art natural language processing.* arXiv preprint arXiv:1910.03771.

Xu, J., Sun, X., Zhang, Z., Zhao, G., & Lin, J. (2019). *Understanding and improving layer normalization*. arXiv preprint arXiv:1911.07013.