

# Safety Report Topic Classification with Transformer-based Data Augmentation

MSc Research Project  
Data Analytics

Jason Payne  
Student ID: 2018543

School of Computing  
National College of Ireland

Supervisor: Muhammad Zahid Iqbal


National College of Ireland  
Project Submission Sheet  
School of Computing



<b>Student Name:</b>	Jason Payne
<b>Student ID:</b>	2018543
<b>Programme:</b>	Data Analytics
<b>Year:</b>	2022
<b>Module:</b>	MSc Research Project
<b>Supervisor:</b>	Muhammad Zahid Iqbal
<b>Submission Due Date:</b>	15/12/2022
<b>Project Title:</b>	Safety Report Topic Classification with Transformer-based Data Augmentation
<b>Word Count:</b>	6,582
<b>Page Count:</b>	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

<b>Signature:</b>	
<b>Date:</b>	13th December 2022

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

Attach a completed copy of this sheet to each project (including multiple copies).	<input type="checkbox"/>
<b>Attach a Moodle submission receipt of the online project submission</b> , to each project (including multiple copies).	<input type="checkbox"/>
<b>You must ensure that you retain a HARD COPY of the project</b> , both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

<b>Office Use Only</b>	
Signature:	
Date:	
Penalty Applied (if applicable):	

# Safety Report Topic Classification with Transformer-based Data Augmentation

Jason Payne  
2018543

## Abstract

Safety Leading Indicators (SLIs) are incident traits, sometimes referred to as weak signals, that, when tracked, enable organisations to proactively plan actions to mitigate significant incident occurrences. This research presents an implementation method for SLIs based on topic classification of safety reports. Rather than imposing a mandatory reporting format, indirect implementation based on text/content analysis significantly reduces implementation complexity and potential for KPI exploitation/manipulation. The method works in low and unlabelled data regimes and is independent of reporting systems, formats and taxonomies. A new multi-label rule-based approach was developed to assign crafted SLI categories to unlabeled safety reports. This labelled data was then used to fine-tune pre-trained Language Models (LMs) for advanced Transformer-based Data Augmentation (TrDA). TrDA was combined with conventional text augmentation techniques to train performant supervised topic classifiers using Bidirectional LSTM (Bi-LSTM) models. The Bi-LSTM models were shown to outperform the upstream rule-based methods on new/unseen data. The proposed methodology is organisation and process agnostic, and the solution is practically deployable via commonly available cloud services.

## 1 Introduction

In the construction industry, safety/incident reports are formal records of HSE-related accidents, incidents or near-misses in the workplace. Reports aim to capture the pertinent aspects of incidents so that root cause analysis can be performed and the organisation's safety performance can be documented per relevant regulatory requirements. Reports are collected digitally, either using proprietary software applications or spreadsheets. While format varies, and no standard categories are applied, common elements of each report include incident contextual information plus short and long text descriptions.

Due to regulatory requirements, organisations put significant effort into collecting and reporting on these data. However, much of the reporting revolves around lagging indicator statistics (Oswald, 2020; Xu et al., 2023). For predictive risk modelling, patterns must be detected that correlate with the risk of particular types of incidents occurring. For example, construction sites recording regular incidents involving site compliance or bad practice issues are likely at a higher risk of a major incident occurring. Site compliance or practice issues are examples of 'weak signals' or 'safety leading indicators'. SLIs are relatively scarce but powerful predictors of future safety performance (Xu et al., 2021).

## 1.1 Research Question, Objectives and Contributions

The goal of this project was to develop methods to improve the detection of SLIs in free-text data and in doing so, help tackle the construction industry problem of SLI implementation (Xu et al., 2023). The project’s research question was as follows:

**Research question:** *“To what extent can text-based data augmentation improve topic classification of safety incident reports to enable effective leading indicator detection and improve construction safety.”*

While the work delivers important contributions to stakeholders in the construction industry, it also benefits stakeholders in software and NLP research domains. Contributions include a new rule-based classification method for SLIs, bespoke synonym dictionaries for SLI rule development, a solution methodology for creating performant supervised topic classification models (where contextual understanding is important and data is unlabelled), and fine-tuned language models for SLI narrative generation. These contributions provide stakeholders with a toolkit to detect the occurrence of SLIs in safety reports at a free-text (content) level, hence are agnostic to local processes and systems. The ability to track leading indicators without having to change local reporting systems gives contractors the ability to improve safety performance and circumvent future incidents.

## 2 Related Work

Aligned with the main elements of the research question (Section 1.1), the review of related-work focused on SLI research, supervised topic classification methods and recently published data augmentation techniques for text in low-data regimes.

### 2.1 Safety Leading Indicators

Xu et al. (2021) performed a systematic review of SLIs in the construction industry as part of the Discovering Safety (DS) programme<sup>1</sup>. The research proposes leading indicator categories based on a comprehensive literature review and industry engagement. The categories range from ‘organisation commitment’ to ‘competence’ and are presented alongside quantitative measures and typical attributes. Building on this research, Xu et al. (2023) take the SLI categories and rank them in terms of importance using systematic survey and voting methods. The research also discusses the barriers to implementation of SLIs in the industry. While general discussion is given to various high-level barriers, no clear solutions are proposed to resolve the implementation problem. Quantitative metrics used in the industry are tentatively proposed but with the caveat that they can often become ‘box-ticking’ exercises (Xu et al., 2023). Oswald (2020) makes a similar point, stating that simply counting interaction numbers as KPIs, rather than assessing the quality/content of the interactions, is ineffective. Similar to accident precursors (Tixier et al., 2016a), SLI occurrences are obvious domain concepts for a predictive safety solution. This project proposes the automatic detection of SLIs in routine accident and observation reports as a solution to the implementation problem. In the proposed solution, the content (text) of routine reports can be continually analysed and there is no dependency on implementation of a common industry taxonomy or reporting method.

---

<sup>1</sup><https://www.discoveringsafety.com>

## 2.2 Rule-based Topic Classification

Topic classification is a method used to programmatically label free-text with target categories. Topic classification is either performed using a rule-based system or by machine learning. Rule-based systems are developed manually by programming hand-made rules based on document content and can outperform other classification algorithms (Aggarwal, 2015). Tixier et al. (2016b) present such an approach for accident precursor classification in safety reports. Where accident precursors are typically objects, such as ‘scaffold’ or ‘ladder’. The method uses regular expression patterns in conjunction with a synonym dictionary and word span constraints to record excellent performance measures (above 90%) albeit for a small dataset (2,200 reports) and simple classification challenge. The use of rule-based systems on construction data is also demonstrated by Marucci-Wellman et al. (2017) where NLP rules are used to augment injury classification alongside traditional machine learning methods. Marucci-Wellman et al. (2017) propose that the manual development of rules is essential for classification of ‘noisy’ free-text incident narratives, especially where there is limited data. Liu and Beldona (2021) demonstrate the use of rule-based methods on social media data. The developed method uses regular expression rules incorporating domain-specific ‘intent phrase’ synonyms, and is combined with a ML-based method in final deployment. A limitation of the methodology is that it relies on manual labelling to build and test both rule and ML-based models.

## 2.3 Topic Classification using Machine Learning

Although powerful, rule-based systems have certain undesirable characteristics. They can be difficult to formulate and require consideration of inter-dependencies and inter-play between rule groups. Also, they are rigid and don’t have the ability to generalise and catch semantic nuances and variations. For this reason, supervised topic classification by machine learning methods can be more effective. Baker et al. (2020) present a comparison of advanced machine learning deep learning methods applied to the classification of safety data. Focusing on the accident precursor challenge, the authors compare performance of deep learning RNN and HAN approaches to more traditional TF-IDF with SVM approach. The deep learning methods did not outperform the traditional method for this relatively simple classification challenge. For simple text classification problems, the presence of certain keywords and terms is often sufficient to make an accurate prediction. The classification of SLIs requires a higher level of contextual understanding. For example, PPE is commonly mentioned in accident reports, but not always in the context of non-compliance. Detecting such nuances requires a method with an ability to process contextuality. Jang et al. (2020) propose a Bi-directional Long Short Term Memory (Bi-LSTM) model with an attention-based final layer to perform sentiment analysis on a large internet movie dataset. By concatenating the outputs of two RNNs that pass the information in forward and backward directions, the Bi-LSTM architecture (Graves and Schmidhuber, 2005) is said to be able to process two-way context from an input sequence (Cornegruta et al., 2016). The attention layer (Bahdanau et al., 2014) chooses which features to focus on before making the final classification.

Another advanced method that can be applied to text classification where the target categories are organised in a tree-shaped taxonomy is hierarchical text classification (Stein et al., 2019). As SLI taxonomy is often presented as a hierarchical structure (Xu et al., 2023), the use of a hierarchical method to classify safety data is interesting. The SLI

classification problem is a multi-label problem (Stein et al., 2019) as the text can belong to several SLI categories. As hierarchical classification inherently applies to multi-class challenges, they require adaptation to provide multi-label output, as well as availability of training data with hierarchical categories assigned. Ahadh et al. (2021) also proposes a hierarchical classification strategy to deal with low-data regimes and class imbalance. Ahadh et al. (2021) proposes a three-step methodology of keyword extraction, topic modelling and hierarchical classification. While the method provides good accuracy for its referenced datasets, over-reliance on keyword distribution for classification is a limitation when applied to more demanding use-cases such as SLI detection.

## 2.4 Data Augmentation

In other domains, data augmentation is commonly applied to improve model performance where data is limited (Shorten and Khoshgoftaar, 2019; Xie et al., 2017). However, in NLP, it is a new and emerging area of research. Wei and Zou (2019) present a set of simple data augmentation techniques based on random synonym replacement and word insertion, swapping and deletion. The research demonstrates that simple augmentation methods can improve classification performance of RNNs for smaller datasets. Although relatively new to text classification, augmentation methods such as random deletion have a proven track record in image classification (Zhong et al., 2020). Building Wei and Zou (2019), Karimi et al. (2021) modify how random operations such as random swapping and deletion are applied to deliver marginally improved performance on test data. Back-translation (Sennrich et al., 2016) is another common text augmentation technique where text is translated to an intermediate language and then translated back to create a modified version to augment the base dataset. Kumar et al. (2020) provide a comparison of standard text data augmentation techniques to an LM-based method for three common LMs (namely BERT, GPT-2 and seq2seq) and three NLP tasks (namely sentiment, intent and question classification). Focusing on low-data regime tasks, the proposed LM-based augmentation involves prepending class labels to training sentence strings. For GPT-2 text generation, the research recommends providing additional context via prompts to assist with preservation of label information. The research concludes that the use of DA improves classification performance in the low-data regime setting (Kumar et al., 2020).

## 2.5 Summary of Findings, Identified Gaps

Table 1 provides a summary of focus research areas with related work, identified gaps and justification for new research. Concerning SLIs, while there is clear industry direction (Xu et al., 2023) on ‘what’ to track, few practical solutions exist or have been put forward for ‘how’ to track SLIs. Effectively tracking SLIs in an organisation or across the construction industry is a complex and challenging task. Implementing a common (enforced) industry-standard taxonomy would be massively challenging (due to its impact on reporting systems) and likely ineffective (due to inherent KPI culture flaws). As per Goodhart’s law (Strathern, 1997), “when a measure becomes a target, it ceases to be a good measure”. Tracking SLIs indirectly through content analysis and not directly through ‘tick-box’ reporting (Oswald, 2020) is proposed in this research project as a more effective approach. However, tracking SLIs indirectly requires effective topic classification of safety reports and, as such, answers to the project’s research question (Section 1.1).

Table 1: Summary of Related-work Findings and Identified Gaps.

Focus area	Main research	Gaps & new research justification
Safety leading indicators (SLIs)	Xu et al. (2021), Xu et al. (2023), Oswald (2020)	<b>Implementation challenge:</b> No practical organisationally agnostic method exists for implementing the proposed SLI taxonomy.
Supervised topic classification & rule-based methods	Tixier et al. (2016b), Baker et al. (2020), Liu and Beldona (2021), Cornegruta et al. (2016)	<b>Low-data &amp; labelled data availability:</b> Potential to enhance existing rule-based methods and combine with advanced ML-methods to create a robust classification solution.
Augmentation for text datasets	Wei and Zou (2019), Kumar et al. (2020)	<b>Unified implementation methodology:</b> Disparate research presented for augmentation methods. Potential to integrate in a unified solution pipeline to improve SLI detection.

### 3 Methodology

Figure 1 presents a process flow diagram for the main steps of the project’s research methodology. The methodology was constructed to enable the training of topic classification models where data is unlabelled and scarce. The methodology combines a Manual Rule-book (MRB) approach with transformer LMs and deep learning classification.

#### 3.1 Data Sourcing and Preparation

Data sourced covered both organisation and regulatory-level reporting. Organisation-level reporting contains incident reports for all severity levels. Regulatory-level data captures data from diverse organisations/sectors and is publicly available. Regulatory data tends to be more succinct than organisation-level data. Organisation-level data tends to be diverse, noisy and can have significant variance in word count. Organisation-level data were provided by a large multi-national energy services company specialising in the design, construction, operation and maintenance of energy infrastructure. Regulatory data was provided by OSHA<sup>2</sup> and is compiled from US construction (severe) accidents. Data from each source were combined to create one dataset. Unwanted data fields such as assigned categories and ratings were dropped. Each row of the combined dataset contained three fields, namely incident ID, text description and data source. The text description was constructed by combining incident title with long description.

#### 3.2 MRB Data Labelling

The first methodology step addressed the problem of unlabelled data. The multi-label classification methodology developed for this project is an enhanced ‘rule-book’ methodology that builds on the basic approach of Tixier et al. (2016b) with a specific focus on SLI detection (rather than accident precursors). The basic premise of the method proposed by Tixier et al. (2016b) was to systematically use synonym dictionaries, regular expression patterns and word span constraints to detect categories of interest. For this project, the methodology was enhanced by implementing ‘keyword-in-context’ (KWIC) searching and context splitting. The KWIC (Luhn, 1960) methodology initialises by searching

<sup>2</sup><https://www.osha.gov>

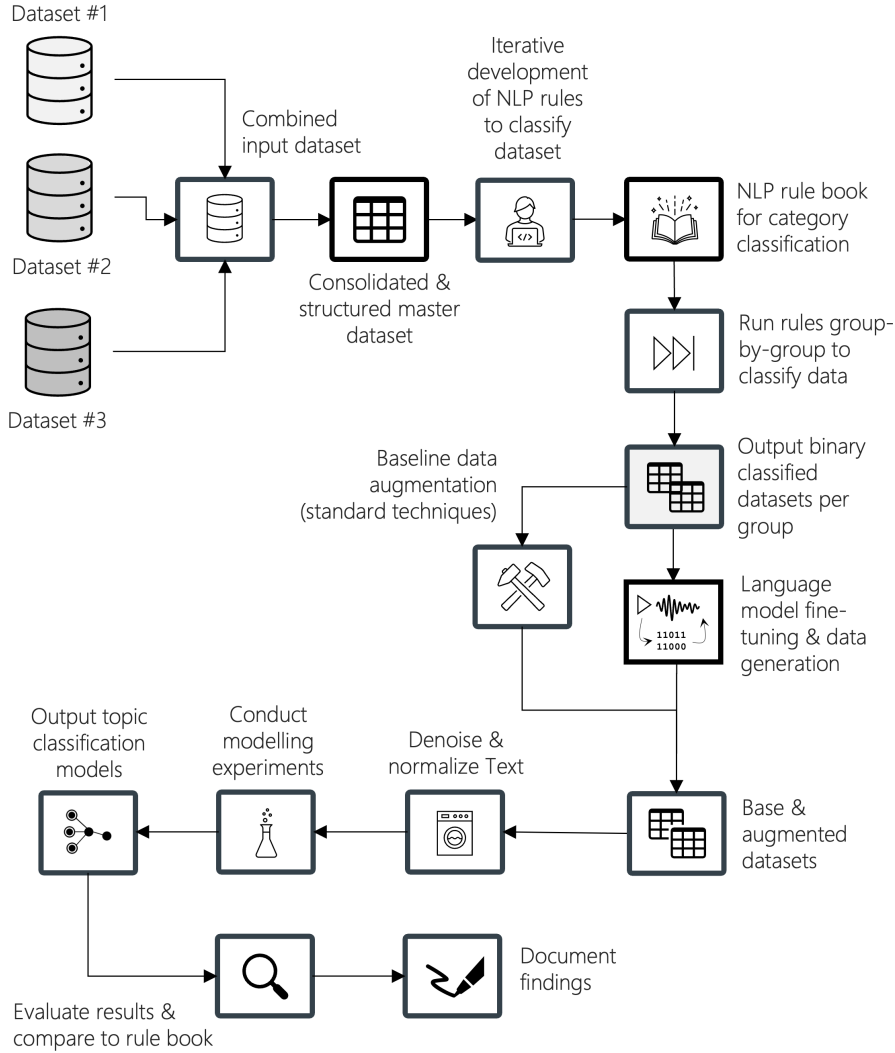


Figure 1: Process Flow for Topic Classification using Data Augmentation.

documents for keywords associated with a particular SLI. If the keyword is found, the context is extracted based on a defined span (i.e., a number of words before and after the keyword). Context is extracted in three parts (i.e., pre-, post- and all-context) so that rules can be constructed to act on the contextual part that is most effective. Contextual splitting significantly reduced the effort required to reach the required classification coverage. The MRB method also allows for the definition of wildcard-type rules where no keyword is required. In the wild-card rule, no word search is performed, each rule pattern is simply applied to each sentence.

### 3.3 Data Augmentation

Data augmentation is used to artificially increase the volume of SLI category training data. Baseline Data Augmentation (BDA) and Transformer-based Data Augmentation (TrDA) methods were used (independently and combined). BDA is an implementation of standard text augmentation techniques based on random selection. These included word insertion, deletion, swapping, and synonym replacement (Wei and Zou, 2019). TrDA used



fine-tuned decoder models and topic-specific prompts to fabricate new incident reports. Prompts were generated using a rule-based approach combined with the SLI synonym dictionary. See Section 5 for specific details of method implementation.

### 3.4 Supervised Topic Classification

Binary supervised topic classifiers were created for five SLI ‘focus’ categories. Focus categories were deliberately selected to challenge model performance. The main challenges were class imbalance and a requirement for contextual processing (not just keyword detection). Supervised topic classification models were selected on the hypothesis that ML-based models would generalise better and be more performant than the MRB. Deep learning methods were selected as the most suitable methods due to the nature and complexity of the classification challenge. However, other more straightforward classifications methods, such as gradient-boosted decision trees were also assessed.

### 3.5 Performance Evaluation

Performance evaluation was carried out for each main step as follows:

**Rule-book coverage and accuracy:** Performance targets adopted for the MRB method are summarised in Table 2. Coverage is defined as the percentage of accident reports with at least one label assigned. Accuracy was assessed by manual review of randomly selected samples of 100 reports. Table 12 presents scoring logic.

**Supervised topic classification:** Supervised classification performance was assessed using the standard metrics of precision, recall and F1-score. Because there was no manually classified ‘ground truth’ for the dataset, the rule-book classifications were used for base performance measure calculation. However, these metrics were then adjusted (where practical) by manual review and false positive correction.

Table 2: MRB Performance Targets.

Measure	Target
Coverage	Greater than 70% of documents with at least one label assigned.
Accuracy	Less than 5% of sampled documents scored as ‘poor’.

## 4 Design Specification

### 4.1 Rule-based Topic Classification

The MRB method combines standard methods (Tixier et al., 2016b) with a bespoke search algorithm based on KWIC (Luhn, 1960). The KWIC-based algorithm initiates with a keyword search. If the keyword is found, contextual splitting is performed on the keyword using a predefined word span (context length). The three contextual elements are extracted, and tailored rules along with void checks are applied to label categories.

## 4.2 Augmentation for Imbalanced Datasets

Data augmentation is used to artificially increase training examples for minority classes. A combination of conventional techniques (Wei and Zou, 2019) and advanced transformer-based techniques were adopted for this project.

## 4.3 Language Models

The pre-trained LM selected for this project was the auto-regressive Generative Pre-trained Transformer 2 (GPT-2) model (Radford et al., 2019). The GPT architecture implements a transformer model (Vaswani et al., 2017) which uses ‘attention’ (Bahdanau et al., 2014) to focus model training on input text segments that it predicts to be the most important. GPT-2 was selected over larger and more recent LMs due to its open source availability (via Hugging face<sup>3</sup>), manageable size and proven NLP track-record.

## 4.4 Embedding Model

As a Bi-LSTM model was pre-selected for the supervised learning task, a relatively simple embedding model was chosen to convert incident report words into numerical representations and capture word similarity. A more sophisticated embedding model, such as one considering contextuality, was not deemed necessary on the premise that the Bi-LSTM model would be able to learn the contextual nuisances of the topic-specific categories. Word embeddings were created using the 6-billion token (50-dimension) database of pre-trained vectors created using Glove (Pennington et al., 2014), made available by Stanford University<sup>4</sup>. GloVe is a global log-bilinear regression model used for unsupervised learning of vector-based representations of words (Pennington et al., 2014). The pre-trained database contains a dictionary of words and their 50d vectors. The narrow 50d version was selected to minimise over-fitting and exclude unnecessary/noisy words.

## 4.5 Bidirectional LSTM (Bi-LSTM)

A multi-layer Bi-LSTM model was selected for the binary-classification task. The LSTM model architecture has a special gate-type structure that helps capture both short and long term dependencies in sequence inputs. The Bi-LSTM is an adaption of the base LSTM model (Graves and Schmidhuber, 2005) in that it comprises both forward and backward LSTM layers. The multi-layer Bi-LSTM has the ability to learn the forward and backward relationships of a sequence and is specifically well suited to time-series and other sequential data such as text (Zan et al., 2019). The multi-layer architecture is illustrated in Figure 2. The main elements of the chosen architecture are the input embedding layer, multiple Bi-LSTM layers, a fully connected dense layer and a softmax classifier. The input embedding layer mutates prepared input into embeddings using the pre-trained GloVe vector dictionary. The multi-layer Bi-LSTM block is used to learn features from the embedding sequences. The fully connected dense layer concatenates the last forward and backward LSTM outputs for input to the softmax classification layer.

---

<sup>3</sup><https://huggingface.co>

<sup>4</sup><https://nlp.stanford.edu/projects/glove/>

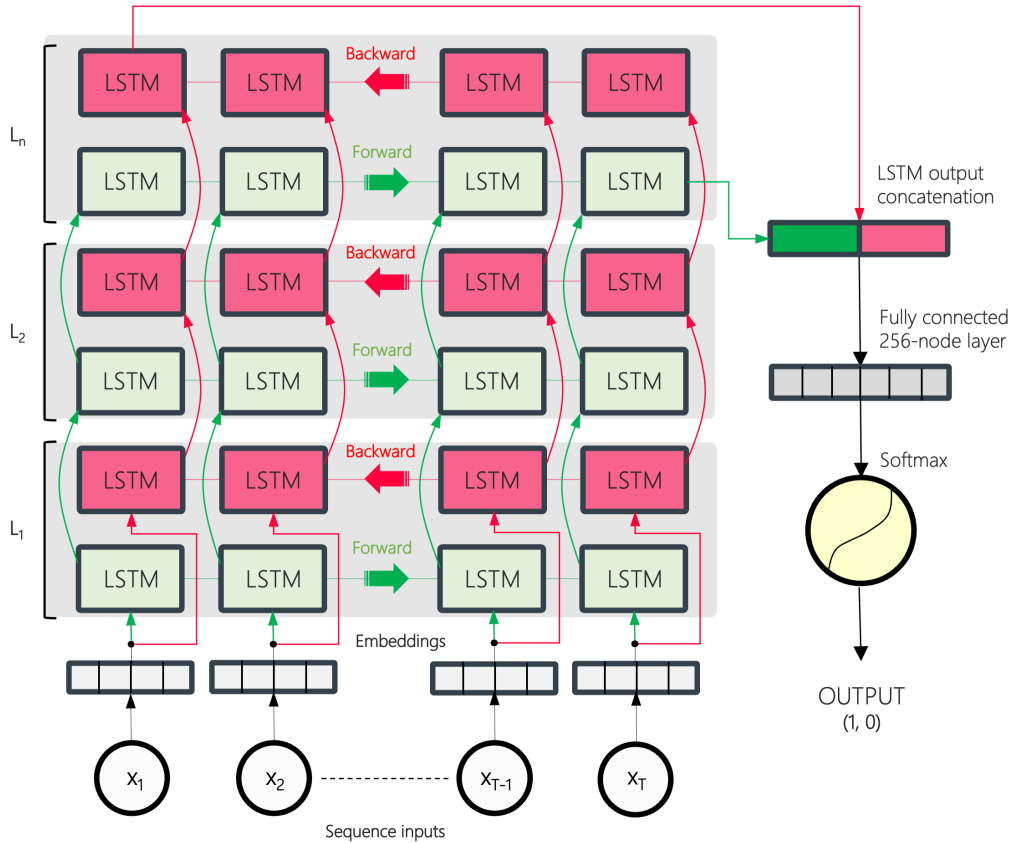


Figure 2: Multi-layer Bi-LSTM Architecture. Layer number is denoted by  $L$  (three are shown for illustration purposes). Sequence inputs are denoted by  $x$ .

## 5 Implementation

### 5.1 Languages, Frameworks and Tools

The research was implemented using Python (version 3.7.15). All script-based code was developed using the Visual Studio Code<sup>5</sup> Integrated Development Environment (IDE). Notebook code was developed using Jupyter format. A summary of the main Python packages used is provided in Table 3 and computing resources are outlined in Table 4. Conda<sup>6</sup> was used for initial virtual environment setup and subsequent package management (installation/removal) was performed using pip<sup>7</sup> and tracked using updates to the project's requirements.txt file located in the code repository.

### 5.2 Data Understanding and Exploratory Analysis

Data understanding and exploratory analysis was performed at different stages in the research to determine document characteristics, verify modelling assumptions and confirm intended output from key data processing steps.

<sup>5</sup><https://code.visualstudio.com>

<sup>6</sup><https://docs.conda.io>

<sup>7</sup><https://pypi.org>

Table 3: Summary of Main Python Packages.

Task	Package	Reference
General data wrangling	Pandas	<a href="https://pandas.pydata.org">https://pandas.pydata.org</a>
NLP pre-processing	nlTK	<a href="https://www.nltk.org">https://www.nltk.org</a>
Pre-trained language models	transformers	Wolf et al. (2019)
Text tokenisers	tokenizers	<a href="https://huggingface.co">https://huggingface.co</a>
Deep learning for language models	PyTorch	<a href="https://pytorch.org">https://pytorch.org</a>
Deep learning for classification models	TensorFlow	<a href="https://www.tensorflow.org">https://www.tensorflow.org</a>

Table 4: Summary of Computing Resources.

Task	Compute Resources
Rule-book development & output evaluation	Windows personal computer with 16GB RAM and Intel Core i7 CPU
Language model fine-tuning & fake text generation	Google Colab <sup>a</sup> ( <a href="https://colab.research.google.com">https://colab.research.google.com</a> ) with Tesla T4 GPU (2560 CUDA Cores)
Deep learning model development	Google Colab <sup>a</sup> with Intel Xeon CPU @ 2.20GHz

<sup>1</sup><https://colab.research.google.co>

### 5.2.1 Data Summary.

Table 5 provides a breakdown of the project’s datasets. Data was sourced from both private company and public sources (Section 3.1) and a small number of manually created reports were created to assist rule development for the ‘competency’ SLI.

Table 5: Dataset Summary.

Source	Records	Description
Private company incidents (ORGP <sup>a</sup> )	27,158	Batch #1 HSE incident reports <sup>b</sup> .
OSHA published incidents	66,699	Severe injury/accident descriptions.
Manually generated incidents	26	Fabricated short narratives.
Private company incidents (ORGP <sup>a</sup> )	2,429	Batch #2 HSE incident reports <sup>c</sup> .

<sup>a</sup> ORGP is a generic reference used for the private company datasets.

<sup>b</sup> Batch #1 was used in all model/method development, training and testing.

<sup>c</sup> Batch #2 was received at the end of the project and was used only for method comparison.

### 5.2.2 Data Understanding

Data understanding and exploratory analysis was performed before model development and after MRB labelling to determine document characteristics and sense-check labelling effectiveness. Figure 3 provides histograms of word counts for incident reports in the two datasets. OSHA incident reports are shorter and have less variation in word count compared to the ORGP reports. This is because OSHA are regulatory reports hence tend to be more succinct and regular than the private ORGP dataset. The private ORGP

dataset contains all incident types and severity levels as reported by the organisation, hence have more variation. 95% of documents in the combined dataset have less than 309 words. Word count statistics are important as they influence GPT-2 model size selection, e.g., the medium GPT-2 model can handle sequences of 1,024 consecutive tokens.

As described in Section 3, labelled datasets for each of the focus SLI categories were created using the MRB method. Word cloud visualisations were created for each data subset as a simple screening check of the effectiveness of the MRB labelling

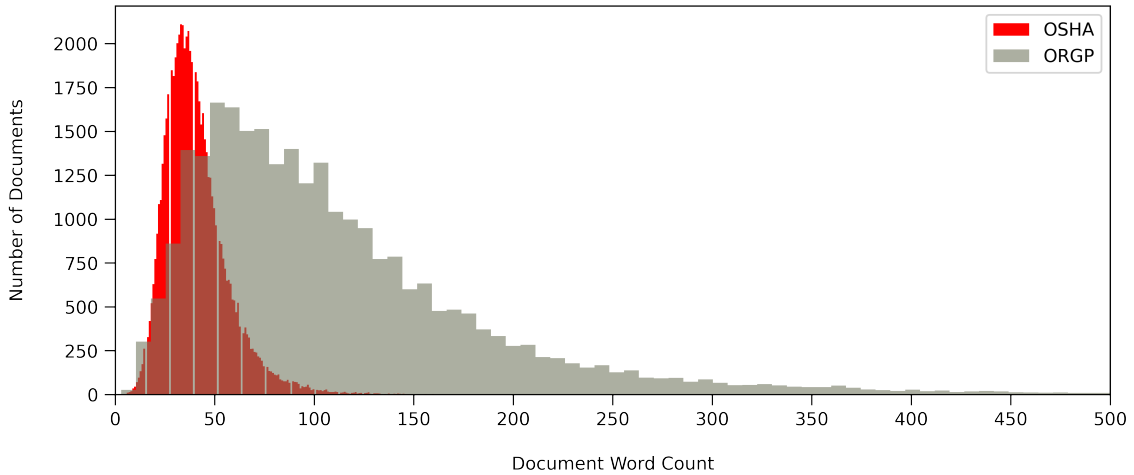


Figure 3: Distribution of document word counts (in each primary dataset). OSHA documents tend to be shorter documents with less variation in word count.

## 5.3 Data Preparation

### 5.3.1 General

The main data preparation/handling steps are summarised in the following sub-sections. Note that test data were fully separated in the development pipeline to avoid leakage. Only training examples of focus categories were used as input to data augmentation.

### 5.3.2 Data Transformation

Before data were used in model training they were transformed using standard text denoising and normalisation techniques. Denoising involved removal of stop sentences, HTML and contradictions. Normalisation involved converting text to lower case, removal of punctuation and non-ASCII characters and replacement of numbers with words. During data preparation, experiments were performed with other denoising and normalisation techniques such as lemmatisation, stemming and Americanisation (i.e., harmonising British and American spelling). These were not found to improve model performance and as such were not carried forward to the final model. Also, stop words were not removed as removing them could impact report semantics and change the classification.

### 5.3.3 Tokenisation and Padding

Tokenisation was used in both Bi-LSTM development and LM fine-tuning as follows:

**Bi-LSTM development:** A standard keras<sup>8</sup> tokenizer was created using the training data. This tokenizer was used to convert training and test data into sequences of integers up to a maximum sequence length of 500. The integer sequences were then padded out with zeroes (where applicable) so that all sequences had 500 entries. The maximum sequence length of 500 was selected based on data understanding (Section 5.2.2).

**Language model fine-tuning:** Each string in the training ‘trues’ data were encoded by the GPT2 tokenizer (from the tokenizer library) to a list of numerical values (one value per word). As GPT-2 model version used for the project could only handle up to 1,024 tokens, some trimming of report word-count was required to meet the model token limit.

### 5.3.4 Embedding Sequences

As stated in Section 4.4 an embedding dictionary was created using pre-trained vectors from GloVe<sup>9</sup>. This dictionary was used to convert the tokenised word sequences into sequences of 50d vector embeddings (as part of the embedding layer of the Bi-LSTM). Vectors for words not found in the embedding dictionary were set to all zeroes.

## 5.4 Data Augmentation

### 5.4.1 General

Data augmentation was performed using BDA (Wei and Zou, 2019) and TrDA techniques. In each technique, SLI focus data from the training ‘trues’ data were used to create approximately 8,000 modified (BDA) and new/fake (TrDA) training observations.

### 5.4.2 Base-line Data Augmentation (BDA)

The BDA method implemented uses a pipeline of the following augmentation techniques based on random selection (Wei and Zou, 2019). In each technique, the integer  $n$  is calculated as 10% of the word count in the document. The target number of modified reports (8,000) was achieved by running the BDA pipeline  $m$  times (calculated based on the number of available reports for the focus category).

**Random Insertion:** Randomly insert  $n$  words into the document. Inserted words are generated from randomly selecting a synonym from a randomly selected word.

**Random Deletion:** Step through each word in the document drawing a sample from a uniform distribution (with interval bounds of 0 and 1). If the drawn sample is less than 0.1, randomly delete the word from the document.

**Random Synonym Replacement:** Randomly select and replace  $n$  words (that are not stop words) with synonyms randomly selected from WordNet (Miller, 1998).

**Random Swap:** Randomly swap two words  $n$  times in each document.

### 5.4.3 Transformer-based Data Augmentation (TrDA)

TrDA used fine-tuned LMs and category specific prompts to fabricate new incident narratives. The GPT-2 (small) model from the huggingface transformer (Wolf et al., 2019)

---

<sup>8</sup><https://keras.io>

<sup>9</sup><https://nlp.stanford.edu/projects/glove/>

package was used for all TrDA models with default training parameters. For text fabrication, Top-k sampling (Fan et al., 2018) was used in combination with Top-p (Holtzman et al., 2019) sampling. According to huggingface documentation (Tunstall et al., 2022), combining the two methods can reduce occurrence of very low ranked words while retaining a level of dynamic selection. Parameter values for  $Top-k = 0.50$ , and  $Top-p = 0.95$  were adopted (Tunstall et al., 2022). LM prompts were created in a similar way to MRB rules using the same synonym dictionary. An example prompt rule (for ‘line strike’) is ‘{vehicleexcavate} {vehiclestruck} {linestruck}’. This prompt expands based on each permutation of the ‘excavation vehicle’, ‘vehicle striking’ and ‘line’ synonyms.

## 5.5 MRB Method

As stated in Section 3.2, a target minimum percentage (70%) of the accident reports were labelled by iterative rule development. This was implemented by randomly sampling 100 reports, crafting regular expression rules and aligning to categories until target coverage was achieved. Once target coverage was achieved (for the sample), a new random sample was selected and the rule development process was continued. The process was stopped when coverage for new randomly drawn samples was consistently above 70%.

Each SLI category had several rules, and the most common rule type was the KWIC-based rule. The implemented method used up to two synonyms per rule. Experimentation with up to three synonyms was performed but found to significantly increase classification time without significantly improving performance. A selection of example rules is given in Table 6. Word/term references enclosed in curly braces denote synonyms taken from the project’s SLI synonym dictionary. Example synonyms are presented in Table 7.

Table 6: Example MRB Rules<sup>a,b</sup>.

Category	Keyword	Rule	Context <sup>c</sup>	Voids
Line strike	strike	{linestruck}	Post	{head}
PPE non-conformance	-	{worker}*.not wear*.{ppe}	All	-
hydraulic fluid or oil leak	leak.*	{fuel}	Pre	-
mechanical or equipment issue	-	failure*{mechcomponent}	All	-

<sup>a</sup> The complete rule-book is compiled in CSV file format.

<sup>b</sup> Wildcard token .\* denotes any combination of characters is possible.

<sup>c</sup> Denotes the contextual part that the rule applies to.

Table 7: Example Synonyms.

Reference	Dictionary examples
worker	apprentice, contractor, cleaner, colleague, employee, ..., woman
ppe	gloves, glasses, goggles, ppe, ..., personal protective equipment
permit	good practice, normal practice, site practice, ..., process
hurt	abrasion, amputat.*, bang.*, break.*, ..., trauma
weather	flood.*, frost, gust, lightning, rain.*, .., wind

## 5.6 Supervised Topic Classification

To assess augmentation performance, four deep learning classification models were created for each focus category. The first was trained without augmentation, the second and third were trained using each data augmentation technique (BDA and TrDA) and the fourth was trained with both augmentation techniques. Model hyper-parameters are summarised in Table 8. Bi-LSTM architecture is summarised in Table 9 and visualised in Figure 2.

Table 8: Model Parameters.

Parameter	Value
Number of epochs	3
Batch size	128
Learning rate	0.001 (default)
Hidden layers	4x bidirectional LSTM layers with recurrent dropout (0.2)

Table 9: Bi-LSTM Architecture.

No.	Layer	Shape	Notes
1	Embedding	500 x 50	Embedding dictionary created from GloVe
2, 4, 6	Bidirectional LSTM	500 x 64	With return sequences & recurrent dropout (0.2)
3, 5, 7	Dropout	500 x 64	Dropout rate = 0.5
8	Bidirectional LSTM	64	With recurrent dropout (0.2)
9	Dropout	64	Dropout rate = 0.5
10	Dense	256	ReLU activation function
11	Dense (output)	2	Softmax activation function

## 5.7 Traceability

Scoring scripts were used to manually score model output. These scripts created output files that were saved with unique filenames to enable independent verification. The files contain raw data, modelling output and the score assigned by manual review.

# 6 Results and Evaluation

## 6.1 MRB Classification Counts

Table 10 and Table 11 summarise classification counts for SLI and ‘general’ categories respectively (with categories with less than 0.2% omitted). With the exception of ‘working at height’ and ‘dropped objects’, each SLI category relates to less than 1% of total reports (less than 800 reports per category). Counts for general categories in Table 11 are in line with expected values. One third (33%) of reports are labelled as ‘hand or arm injury’ and approximately 20% of incidents labelled as either ‘foot or leg injury’ or ‘slips and trips’. 38% of incidents in the raw OSHA<sup>10</sup> dataset have affected body part labelled as

<sup>10</sup><https://www.osha.gov/severeinjury>



being an arm or hand body part. [HSE \(2021\)](#) report ‘slips, trips or falls on same level’ as representing 33% of all non-fatal accidents in 2021/2021. Note that both OSHA and HSE statistics are based on regulatory requirements for reportable incidents, the ORGP (private) dataset includes all incident types hence perfect alignment was not expected.

Table 10: SLI Categories & Counts - MRB Classification.

Category <sup>a</sup>	Reports classified	% of total <sup>b</sup>
Fall from or working at height issue	9,009	9.6
Dropped object or material	5,203	5.5
Hazardous materials or work	759	0.8
Line strike	696	0.7
Hydraulic fluid or oil leak	583	0.6
Mechanical or equipment issue	504	0.5
Site compliance or practice issue	400	0.4
Near miss	290	0.3
Line of fire	267	0.3
Fuel spill or leak	226	0.2
Environmental leak or issue	168	0.2
PPE non-compliance	144	0.2

<sup>a</sup> Highlighting denotes focus categories, i.e. those selected for supervised model training.

<sup>b</sup> Total number of rows in the ‘raw’ dataset (i.e., 93,857).

## 6.2 MRB Coverage

To quantify MRB coverage, 30 experiments were conducted. Each experiment calculated coverage for 100 reports selected by random proportionate sampling. This method avoided running all rules on the entire dataset, which would have required significant stable compute time. Figure 4 presents percentage coverage for each experiment. Target coverage at the outset of the project was 70%, median recorded coverage was 88%.

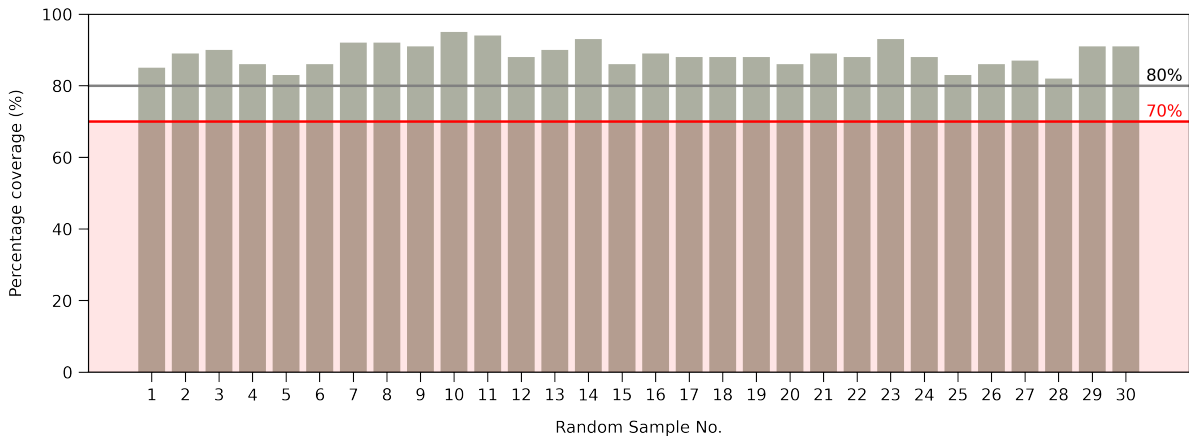


Figure 4: MRB Percentage Coverage. 30 samples were selected by random proportionate sampling and classified by the rule-book method. Target minimum coverage was 70%.

Table 11: General Categories &amp; Counts - MRB Classification.

Category	Classified reports <sup>a</sup>	% of total <sup>b</sup>
Hand or arm injury	30,468	32.5
Foot or leg injury	18,098	19.3
Slips and trips	18,070	19.3
Vehicle incident general	8,812	9.4
Lifting or moving loads	8,402	9.0
Head or face injury	7,720	8.2
General injury	6,252	7.0
Covid	5,877	6.3
Back injury	4,500	4.8
General illness or health issue	1,992	2.1
Weather related issue	1,226	1.3
Eye injury or irritation	1,207	1.3
Burns	1,016	1.1
Electric shock or issue	696	0.7
Bites (animals or insects)	516	0.5
Fire	511	0.5
Knife incident	482	0.5
Vehicle Incident involving animal	414	0.4

<sup>a</sup> Classified denotes a rule-book ‘hit’ for the category in a report/document

<sup>b</sup> Total number of rows in the ‘raw’ dataset (i.e., 93,857)

### 6.3 MRB Label Quality Scoring

The quality of MRB labelling (for the final model) was assessed by conducting five separate scoring experiments. Each experiment involved sampling 100 labelled reports and manually assigning a score of ‘good’, ‘fair’ or ‘bad’, using the scoring logic in Table 12. Sampling was conducted to yield approximately a 50:50 split of OSHA and ORGP incidents in the sample. Disproportionate sampling was performed to provide a greater test as the longer/less succinct ORGP reports pose a greater test for the model. Results of the scoring experiments are provided in Table 13. All experiments yield less than 5% of reports with ‘bad’ labels. The average result for ‘good’ labels is 76%. As discussed (Section 5.5), rules were created in an iterative manner until target coverage and quality was consistently exceeded. Once exceeded, the final evaluation was performed.

Table 12: MRB Labelling - Scoring Definitions.

Score	Definition
Good	All labels are accurate and no significant category is missing.
Fair	At least one significant category is assigned but some applicable categories could be missing or assigned incorrectly.
Bad	None of the assigned categories are applicable to the accident description.

Table 13: MRB Labelling - Scoring<sup>a</sup> of Samples.

No.	Sample Ref. <sup>b</sup>	Sample Size <sup>c</sup>	% Good	% Fair	% Bad
1	220801172342	100	76	22	2
2	220806193242	100	74	23	3
3	220806204123	100	79	17	4
4	220908162210	100	76	20	4
5	220908222547	100	73	27	0

<sup>a</sup> See Section 6.1 for definitions of ‘good’, ‘fair’ and ‘bad’.

<sup>b</sup> Filename references are for experiment output files containing all labels and manual scoring.

<sup>c</sup> Actual origin sample size is larger to yield greater than 100 classified docs/reports.

## 6.4 Supervised Topic Classification Performance

Table 14 presents a performance summary for the Bi-LSTM models. For each category, results are presented for four model types, covering each augmentation level. The results demonstrate that the Bi-LSTM model is a good choice for SLI detection and that data augmentation techniques can significantly improve performance. Combining augmentation techniques yielded the best performance, with each model recording high recall against rule-book hits (0.7 to 0.9) and adjusted precision (0.6 to 0.8). Based on candidate’s research for this project, no published performance measures exist for SLI classification in safety reports. Elsewhere in the safety domain, Tixier et al. (2016b) reported classification accuracies above 0.9 for safety precursors, Marucci-Wellman et al. (2017) reported accuracies of 0.8 to 0.9 for injury and illness events, and Baker et al. (2020) reported accuracies of 0.6 to 0.8 for standard incident report categories. Considering the significant class imbalances associated with the chosen SLI categories, and the complexity of the detection problem compared to previous classification work in the domain, the performance measures achieved in this research are considered to be comparably strong.

Model experiments for the ‘hydraulic fluid or oil leak’ category considering other model types yielded poor performance (see Table 15). A gradient-boosted tree using CatBoost (Dorogush et al., 2018) yielded recall and precision accuracies of 0.58 and 0.02 respectively. Logistic regression yielded similar measures to CatBoost (recall 0.53 and precision 0.01). For this scenario, Bi-LSTM delivered recall of 0.88 and precision of 0.52.

Table 16 presents a performance comparison for MRB and Bi-LSTM methods on a ‘new’ dataset of approximately 2,500 records. This dataset was received at the end of the project and was unseen by both MRB and Bi-LSTM methods. The method comparison results indicate that the Bi-LSTM performs better than the MRB on ‘new’ data. The Bi-LSTM detects more SLIs while maintaining low false positive rates and performs particularly well for complex and rare categories such as ‘site compliance and practice issues. Increased detection counts were recorded for four of the five categories.

## 6.5 Discussion

The findings from this research project demonstrate that Bi-LSTM models can be trained on relatively small data volumes to effectively label SLIs in construction incident reports. The classification methodology is suitable for any domain with unlabelled text, and traditional ‘bag-of-words’ type approaches are ineffective. It is particularly useful in use cases where large volumes of data must be labelled according to a new labelling taxonomy, e.g.

Table 14: Supervised Topic Classification - Bi-LSTM Performance Measures<sup>a</sup>.

Focus group/model	Precision <sup>a</sup>		New	Recall	F1
	Base	Adj.	Finds	Base	Adj.
<b>Hydraulic fluid or oil leak (0.62%)<sup>b</sup></b>					
1. Bi-LSTM	0.52	0.76*	43 (40%)	0.88	0.83*
2. Bi-LSTM + BDA	0.65	-	-	0.88	0.87
3. Bi-LSTM + TrDA	0.74	-	-	-	0.71
4. Bi-LSTM + BDA + TrDA	0.53	0.73*	36 (34%)	0.90	0.80*
<b>Site compliance or practice issue (0.43%)<sup>b</sup></b>					
1. Bi-LSTM	0.51	0.84*	16 (20%)	0.31	0.45*
2. Bi-LSTM + BDA	0.19	-	-	0.49	0.27
3. Bi-LSTM + TrDA	0.35	-	-	0.27	0.30
4. Bi-LSTM + BDA + TrDA	0.21	0.71*	98 (166%)	0.71	0.71*
<b>Mechanical or equipment issue (0.50%)<sup>b</sup></b>					
1. Bi-LSTM	n	n	n	n	n
2. Bi-LSTM + BDA	0.34	-	-	0.33	0.33
3. Bi-LSTM + TrDA	0.14	-	-	0.26	0.15
4. Bi-LSTM + BDA + TrDA	0.54	0.83*	53 (50%)	0.90	0.86*
<b>Line strike (0.35%)<sup>b</sup></b>					
1. Bi-LSTM	0.38	0.90*	32 (49%)	0.35	0.47*
2. Bi-LSTM + BDA	0.42	-	-	0.28	0.34
3. Bi-LSTM + TrDA	0.37	-	-	0.31	0.34
4. Bi-LSTM + BDA + TrDA	0.41	0.81*	58 (94%)	0.95	0.87*
<b>PPE non-compliance (0.15%)<sup>b</sup></b>					
1. Bi-LSTM	n	n	n	n	n
2. Bi-LSTM + BDA	0.09	-	-	0.14	0.11
3. Bi-LSTM + TrDA	0.22	-	-	0.19	0.20
4. Bi-LSTM + BDA + TrDA	0.42	0.59*	11 (34%)	0.84	0.65*

<sup>a</sup> Adjusted measures (denoted by\*) have been corrected by manual review of FPs.

<sup>b</sup> Test data imbalance (i.e., positive count as a percentage of negative count).

Table 15: Supervised Topic Classification - Other Model Experiments<sup>a</sup>.

Focus group/model	Base Precision	Recall
<b>Hydraulic fluid or oil leak (0.62%)</b>		
1. CatBoost	0.02	0.58
2. Logistic Regression	0.01	0.52

<sup>a</sup> No data augmentation or other treatment for data imbalance applied.

when the goals/focus of an organisation change. The research scope included three demanding modelling tasks: MRB generation, language model fine-tuning and deep learning model development. The demands of base model creation meant it was not feasible to spend significant time on architecture and hyper-parameter optimisation. As such, the selections made for this project could be sub-optimal. Also, it was not feasible to perform several model runs (with different random seeds) to account for stochasticity. Only one seed was used, and all results were presented for the same seed. Due to the complexity

Table 16: Topic Classification - Method Comparison (2,429 Unseen Records).

Focus group/method	Finds	Good <sup>a</sup>	Gains <sup>b</sup>	Bad <sup>b</sup>
<b>Hydraulic fluid or oil leak</b>				
1. Rule-book method (RBM)	31	25 (81%)		6 (19%)
2. Bi-LSTM + BDA + TrDA	37	29 (78%)	+4	8 (22%)
<b>Mechanical or equipment issue</b>				
1. Rule-book method (RBM)	11	2 (18%)		9 (82%)
2. Bi-LSTM + BDA + TrDA	41	17 (41%)	+15	24 (59%)
<b>Site compliance or practice issue</b>				
1. Rule-book method (RBM)	6	3 (43%)		4 (57%)
2. Bi-LSTM + BDA + TrDA	21	13 (62%)	+10	8 (38%)
<b>Line strike</b>				
1. Rule-book method (RBM)	8	6 (75%)		2 (25%)
2. Bi-LSTM + BDA + TrDA	21	14 (67%)	+8	7 (33%)
<b>PPE non-compliance</b>				
1. Rule-book method (RBM)	7	5 (71%)		2 (29%)
2. Bi-LSTM + BDA + TrDA	7	4 (57%)	-1	3 (43%)

<sup>a</sup> Number of labels judged to be a correctly or badly assigned by manual review.

<sup>b</sup> Increase in detection count relative to MRB method.

of the various modelling steps, maintaining repeatability, even with a consistent seed, was challenging. Improvements to code structure and abstraction of different code blocks could have improved this.

Rule-book development stopped when coverage metrics (greater than 70%) and quality (less than 5% ‘bad’) were achieved. With more time, adding new rules and enhancing existing ones would improve MRB performance and the bespoke synonym dictionary. Improved rules would generate higher quality (and volume) training data for LM fine-tuning and Bi-LSTM training. Also, no significant time could be given to screening the training ‘trues’ used as input to LM fine-tuning and TC modelling. Filtering out badly labelled documents from the training data could have improved overall performance.

In terms of embeddings for Bi-LSTM input, the 50d GloVe pre-trained vector set was selected for both technical and practical reasons. While the selected model delivered good performance, conducting experiments with a higher-dimension vector set to assess the impact on performance would have been helpful. Alternative embedding models/approaches could also be assessed. The good performance of the Bi-LSTM, compared to the decision tree and logistic regression experiments, is considered to validate the decision to use an embedding model without contextuality.

A high number of augmented data points (8,000) were selected for both augmentation techniques. No sensitivity analyses were performed to determine if this selection was optimal, and the number of augmentations per piece of original text for the BDA techniques was significantly higher than would typically be applied. However, as good performance was observed for the base case approach, and research time was limited, no further experimentation was performed. Also, concerning augmentation techniques, back-translation was not included due to time constraints. Processing time for back-translation of c.93k

reports would take significant time and, as such, was omitted

## 7 Conclusion and Future Work

The goal of this research was to determine the extent to which text-based data augmentation can improve topic classification of safety reports and, in doing so, enable effective SLI detection and improve construction safety. To this end, a three-part solution was developed comprising a new rule-based multi-label classification method, fine-tuned LMs (for enhanced data augmentation) and binary multi-layer Bi-LSTM classification models. The first step involved creating a new rule-based model to tentatively label the construction incident report database according to target SLIs. Five relatively scarce SLI categories were then selected and augmented using random selection techniques and transformer-based augmentation (TrDA). For TrDA, pre-trained GPT-2 language models were fine-tuned for each of the five selected SLI categories. These fine-tuned language models were used with topic-specific prompts, mutated from rules using the SLI synonym dictionary, to create fake incident narratives for each of the five selected SLIs. The final step involved creating binary Bi-LSTM models to detect the presence of the SLI categories in text-based incident reports. The Bi-LSTM models, with augmented training data, were shown to be effective SLI detection models. Also, the Bi-LSTM models generalised well, were more robust than the rule-based method and did not yield high/undesirable false positive rates. The research shows that combining text augmentation techniques improves detection performance significantly. The research delivers a methodology and toolkit (i.e., fine-tuned language models, tailored synonym dictionaries) for SLI detection and adaption to other uses-cases and domains. Due to time constraints and the complexity of the project's constituent parts, experimentation with different architectures and hyper-parameters was limited. Although a performant methodology was developed and demonstrated, some of the selected parameter values could be somewhat sub-optimal.

There is significant potential for the commercialisation of a multi-label classification model for SLIs in the construction industry. The methodology developed is organisation and process agnostic, and could be deployed easily using standard cloud computing services. Construction organisations (or software vendors) could pass text from reports through an API for SLI detection. Classifications could then be used for dashboard reporting or as features for more advanced predictive risk modelling. In terms of improvements, there is scope through using LM techniques such as summarisation and SLI-specific questions/answers for novel feature engineering. Also, implementation of a hybrid classification model and/or enhancement of the rules and synonym dictionaries. In terms of TrDA methodology improvements, experimentation with methods where the hard 'prompt' text occurs intra-narrative could be interesting. Also, the research shows that giving more time to generating high-quality prompts improves augmentation performance.

## Acknowledgement

I would like to thank my tutor Muhammad Zahid Iqbal for his guidance throughout this research project. I would also like to thank my employer for supporting me throughout this MSc programme. Finally, this research project is dedicated to the memory of my mother, who sadly passed this year, *Go Raibh Suaimhneas Síoraí Uirthi*.

## References

- Aggarwal, C. C. (2015). *Data Classification: Algorithms and Applications*, CRC Press.
- Ahadh, A., Binish, G. V. and Srinivasan, R. (2021). Text mining of accident reports using semi-supervised keyword extraction and topic modeling, *Process Safety and Environmental Protection* **155**: -455–465.
- Bahdanau, D., Cho, K. and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate, *arXiv:1409.0473* .
- Baker, H., Hallowell, M. R. and Tixier, A. J.-P. (2020). Automatically learning construction injury precursors from text, *Automation in Construction* **118**: 103145.
- Cornegruta, S., Bakewell, R., Withey, S. and Montana, G. (2016). Modelling radiological language with bidirectional long short-term memory networks, *arXiv preprint arXiv:1609.08409* .
- Dorogush, A. V., Ershov, V. and Gulin, A. (2018). Catboost: gradient boosting with categorical features support, *arXiv preprint arXiv:1810.11363* .
- Fan, A., Lewis, M. and Dauphin, Y. (2018). Hierarchical neural story generation, Vol. 1, pp. 889–898.
- Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures, *Neural networks* **18**(5-6): 602–610.
- Holtzman, A., Buys, J., Du, L., Forbes, M. and Choi, Y. (2019). The curious case of neural text degeneration, Vol. 2540.
- HSE (2021). Kind of accident statistics in great britain 2021.
- Jang, B., Kim, M., Harerimana, G., Kang, S.-u. and Kim, J. W. (2020). Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism, *Applied Sciences* **10**(17): 5841.
- Karimi, A., Rossi, L. and Prati, A. (2021). Aeda: An easier data augmentation technique for text classification, pp. 2748–2754.
- Kumar, V., Choudhary, A. and Cho, E. (2020). Data augmentation using pre-trained transformer models, *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, Association for Computational Linguistics, Suzhou, China, pp. 18–26.
- Liu, Y. and Beldona, S. (2021). Extracting revisit intentions from social media big data: a rule-based classification model, *International Journal of Contemporary Hospitality Management* .
- Luhn, H. P. (1960). Key word-in-context index for technical literature (kwic index), *American documentation* **11**(4): 288–295.

- Marucci-Wellman, H. R., Corns, H. L. and Lehto, M. R. (2017). Classifying injury narratives of large administrative databases for surveillance—a practical approach combining machine learning ensembles and human review, *Accident Analysis & Prevention* **98**: 359–371.
- Miller, G. A. (1998). *WordNet: An electronic lexical database*, MIT press.
- Oswald, D. (2020). Safety indicators: questioning the quantitative dominance, *Construction Management and Economics* **38**(1): 11–17.
- Pennington, J., Socher, R. and Manning, C. D. (2014). Glove: Global vectors for word representation, *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019). Language models are unsupervised multitask learners.
- Sennrich, R., Haddow, B. and Birch, A. (2016). Improving neural machine translation models with monolingual data, *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, Vol. 1, pp. 86–96.
- Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning, *Journal of big data* **6**(1): 1–48.
- Stein, R., Jaques, P. and Valiati, J. (2019). An analysis of hierarchical text classification using word embeddings, *Information Sciences* **471**: 216–232.
- Strathern, M. (1997). ‘improving ratings’: audit in the british university system, *European Review* **5**(3): 305–321.  
**URL:** [https://doi.org/10.1002/\(sici\)1234-981x\(199707\)5:3;305::aid-euro184j3.0.co;2-4](https://doi.org/10.1002/(sici)1234-981x(199707)5:3;305::aid-euro184j3.0.co;2-4)
- Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B. and Bowman, D. (2016a). Application of machine learning to construction injury prediction, *Automation in construction* **69**: pp-102–114.
- Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B. and Bowman, D. (2016b). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports, *Automation in Construction* **62**: 45–56.
- Tunstall, L., Werra, L. v. and Wolf, T. (2022). *Natural language processing with transformers: Building language applications with hugging face*, O’Reilly.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017). Attention is all you need, *Advances in neural information processing systems* **30**.
- Wei, J. and Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks, *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 6382–6388.



- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. et al. (2019). Huggingface’s transformers: State-of-the-art natural language processing, *arXiv preprint arXiv:1910.03771* .
- Xie, Z., Wang, S., Li, J., Lévy, D., Nie, A., Jurafsky, D. and Ng, A. (2017). Data noising as smoothing in neural network language models.
- Xu, J., Cheung, C., Manu, P. and Ejohwomu, O. (2021). Safety leading indicators in construction: A systematic review, *Safety Science* **139**: 105250.
- Xu, J., Cheung, C., Manu, P., Ejohwomu, O. and Too, J. (2023). Implementing safety leading indicators in construction: Toward a proactive approach to safety management, *Safety Science* **157**: 105929.  
**URL:** <https://www.sciencedirect.com/science/article/pii/S09257753522002685>
- Zan, T., Liu, Z., Su, Z., Wang, M., Gao, X. and Chen, D. (2019). Statistical process control with intelligence based on the deep learning model, *Applied Sciences* **10**(1): 308.
- Zhong, Z., Zheng, L., Kang, G., Li, S. and Yang, Y. (2020). Random erasing data augmentation, *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence* pp. 13001–13008.