ISSN 2754-1118 (Online)
ISSN 2753-6599 (Print)

Enhancing Bowtie Hazard Analysis

A Natural Language Processing Approach for Extracting Bowtie Barriers and Their Effectiveness from Occurrence Reports

Jon Ricketts

System Safety Engineer, Christchurch, UK

Abstract

Bowties are a common hazard analysis tool used within aerospace and other safety critical industries, where they have the advantage of clearly depicting relationships between causes, hazards, consequences, and barriers (hazard controls). For a given bowtie, it is useful to understand how these barriers are performing, which is where occurrence reports provide an insight—describing incidents and how effective the barriers were. This paper presents a Natural Language Processing method—Aspect Sentiment Triplet Extraction, adapted to extract barriers from free-text occurrence reports alongside their effectivity and sentiment—a distinct advantage over pure entity extraction methods. The development of a bespoke labelled dataset to train a Bidirectional Encoder Representations from Transformers (BERT) model required extensive safety expertise, where the resulting model output was then qualitatively assessed. The model was ultimately applied to occurrence reports pertaining to lithium battery issues, proving useful in demonstrating key areas where the associated bowtie barriers were operating effectively, ineffectively, or in situations where no barriers were seemingly present. This was especially evident when the output was processed using an unsupervised k-means algorithm to display generalized themes. The method forms a useful tool for safety professionals who are concerned with extracting bowtie barrier information from high quantities of occurrence or incident data while addressing a critical gap in the body of knowledge by going beyond present entity extraction methods.

1 Introduction

This paper presents a safety management method for the automatic extraction of bowtie barriers and associated effectiveness from free-text occurrence reports. The method allows bowtie barriers not only to be extracted *en masse*, but also detail pertaining to their effectivity and an overall score of: positive, negative or neutral. Since it is often difficult to ascertain whether the barriers of a bowtie are operating effectively, the method forms a useful tool for safety professionals who are concerned with extracting bowtie barrier information from high quantities of occurrence or incident data. The method also addresses a critical gap in the body of knowledge by going beyond present, sole entity extraction methods.

Occurrence reports describe situations where near-misses or accidents occurred; therefore they contain valuable information for safety professionals regarding the ongoing assurance of safety (Ricketts et al. 2023). Meanwhile, system safety is typically modelled via a number of common analysis techniques where bowties represent one such method. Bowties are a model; such a model is a result of an abstraction process used to simplify a complex system or problem (Prenninger and Pretschner 2005). Being a predictive model, a bowtie cannot be fully validated, but may be disproved or revised based upon in-service feedback.

A key area of interest is to understand how the bowtie barriers (also described as hazard controls) performed during any given occurrence. If collected over time, this information can provide valuable insight into the ongoing effectiveness of given barriers throughout a system. The results from this can then verify safety and hazard analysis artefacts, providing evidence not only to the effectiveness of barriers but also validating assumptions made during the hazard analysis.

Previous work has sought to classify occurrences against threat pathways of a bowtie (Hughes et al. 2018) or extract causes and consequences from occurrences (Ricketts et al. 2022). As far as the author is aware, occurrence reports have not been used for combined extraction and effectiveness assessment of barriers or other hazard attributes. Therefore, the new method described in this paper demonstrates a possibility for rapid and accurate extraction of safety information, leading to improved safety management.

Occurrence reports provide information from the in-service perspective, providing valuable information to design and production organizations with regards to safety and whether the system is functioning as intended. Where occurrence reviews can be undertaken manually for small, niche systems, a greater challenge is presented for data rich systems. For example, the UK Civil Aviation Authority (CAA) estimate that they receive over 30,000 occurrence reports per year (CAA 2023); it is often unachievable with resource constraints for human analysts to parse this data and capture all relevant knowledge. Fortunately, the development of Natural Language Processing (NLP) provides various solutions to this problem.

NLP brings together the fields of computer science and linguistics. NLP provides solutions to language processing tasks that range from simple rule-based systems through to more sophisticated machine learning and artificial intelligence methods. The area of NLP explored in this paper is that of Aspect Sentiment Triplet Extraction (ASTE), concerned with extracting an aspect (e.g. a barrier) and the associated sentiment towards said aspect (Nazir et al. 2022). The drive for ASTE was originally from the e-commerce domain, where there is a need to parse large quantities of reviews and user experiences on products, building upon sentiment analysis and opinion mining (Liu 2012).

The process presented in this paper describes the creation of a bespoke training dataset and model based upon Mandatory Occurrence Reports submitted to the CAA pertaining to the Carriage of Lithium Batteries, of which there is an operational bowtie[1]. The results of the model solution can then be used to validate the barriers within the bowtie. The clear bounding of the scope to lithium batteries allows the method to be assessed and understood prior to repeating across wider, diverse data.

The carriage of lithium batteries represents a fire hazard onboard aircraft. If a battery has manufacturing defects, or has been mechanically or electrically damaged, this can lead to undesired electrochemical reactions. These can then lead to battery rupture and fire or explosion due to the reaction of released flammable gases from the battery combined with

[1] Found at: https://www.iata.org/contentassets/05e6d8742b00047259bf3a700bc9d42b9/ukca_-bowtie_model_carriage_of_lithium_batteries.pdf

the ambient oxygen in the local atmosphere (Chen et al. 2021). Therefore, an NLP solution that can gather safety data for a bowtie concerned with lithium batteries has the potential to improve safety.

This paper's main contribution and aim is the development and assessment of an ASTE method for extracting and assessing bowtie barriers from occurrence data. Section 2 describes the theoretical background to the areas of research drawn together by the paper. Section 3 describes the method used to create the dataset and model. Section 4 presents the results, both from a metrics perspective and qualitative analysis. Section 5 discusses these results alongside limitations and areas of future work, while Section 6 concludes.

2 Background

The following sub-sections provide an overview of the areas brought together in this paper and associated previous work.

2.1 Hazard Analysis — Bowties

Bowties have been extensively used in the high hazard process industries (de Ruijter and Guldenmund 2015). They have since been incorporated into, and have become commonplace in, other safety critical industries such as aviation, complementing Reason's (2000) Swiss-cheese Model. Hence, due to their widespread use, this paper seeks to integrate an NLP method with a given bowtie to understand if it will be beneficial.

De Ruijter and Guldenmund (2015) describe how bowties have developed from fault trees, event trees, cause consequence diagrams and barrier thinking. Presently, there is not a consistent approach to developing bowties, although the general principle remains the same and is captured within the ISO 31010:2019 standard. Through developing a bowtie, the threats, hazards, and consequences can be identified and mapped into the bowtie format. Once identified, barriers can be implemented with the purpose of preventing the hazard from occurring or reducing the severity of the outcome. This data can then be represented in tabular hazard logs or pictorially through bowties that clearly depict the barriers that prevent the initiating event or accident.

As mentioned in the introduction, a bowtie is an abstraction and is useful as a conceptual model, although they do have some limitations. One problem with bowties is that they can be developed to be too generic; for example, a barrier may be recorded as 'design', however this does not tell the reader what aspects of design are preventing the hazard being realized—admittedly a judgement is required as to what constitutes too little and too much information when constructing bowties, as well as the intended readership. Likewise, a bowtie may not always be intuitive, depending upon the reader's background and perspective.

Once a bowtie has been developed for an in-service system, it is important to understand if the barriers:

- Reflect reality — Are they used as intended within the system and by operators?
- Match assumptions — Were assumptions made by the design organization accurate?
- Are effective — Are the barriers actively preventing accidents from occurring?
- Introduce any unintended consequences — Do they introduce an unsafe element?
- Are known — are there active barriers that are not captured within the bowtie?

Occurrences are a key source of information to answer the queries listed above, being that they allow the operator to anonymously record events relating to the safety of the system.

2.2 Occurrences & Relation to Bowties

The method employed by this paper relies on a dataset of occurrence reports where the free-text of the occurrence typically describes the event and provides indication of the various bowtie components. A dataset with this feature can typically be found in any safety critical industry, however, for the purposes of this paper, CAA Mandatory Occurrence Report (MOR) data was used.

Each MOR is written in English, featuring a headline title then a narrative—that is the focus of this work. The narrative can be written by reporters from a variety of backgrounds, hence there is no set format. They may be verbose or succinct and can feature various acronyms and polysemic terms, which can sometimes pose an issue for pre-trained machine learning models that have not encountered such terms in their training data.

A single occurrence may describe one or more elements of a bowtie, typically describing the consequence from a related cause (if known) and potentially assessing barriers. If many occurrences are viewed together then the likelihood of capturing the entire bowtie or more unforeseen features becomes greater, which is where extractive NLP techniques can play an important role in safety management systems—this is demonstrated in Figure 1.

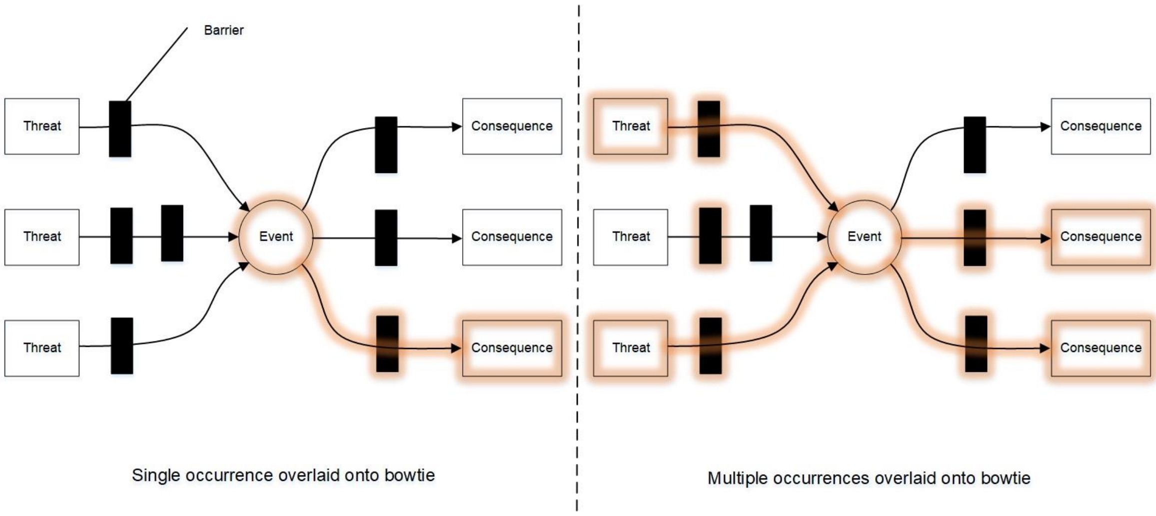

Figure 1~ Single Occurrence vs. Multiple Occurrences Overlaid onto a Bowtie

Work such as that described in this paper could pave the way for occurrence data to be directly connected to safety analysis artefacts, providing timely information on system safety.

2.3 Extracting Barriers and Effectiveness from Free-Text

The third, and main component of this paper is the development of the NLP solution that will be able to satisfy the aim. Where it may be possible to extract barriers as entities and perform sentiment analysis for the whole occurrence, this does not provide the accuracy sought by this paper. Several open-source models excel at entity extraction, or Named Entity Recognition (NER), such as "Universal NER" (Zhou et al. 2023) or "GLiNER" (Zaratiana et al. 2023) that will attempt to extract custom, user defined entities from text.

Both are generalized models that do not always extract a custom entity primarily due to the unique language of occurrences. Also, they are not capable of extracting the associated opinion or sentiment of the entities. A consideration with deploying machine learning methods is how they will interact with the 'language' of occurrences, which, as described by Ricketts *et alia* (2022), can often feature many acronyms and unique, terse, and polysemic terms. Not only may machine learning models need to be trained on these occurrences but adaptation of features may be required, for example, sentiment can be portrayed in a different way to product reviews or social media posts, which typically contain more emotive terms.

It is also important prior to commencing the method, to define an *effective* barrier. A preventative bowtie barrier is deemed "effective" if it performs the intended function when demanded, to the intended standard, and it is capable on its own of preventing a threat from developing into the top event. A mitigation barrier is "effective" if it is capable of either completely mitigating the consequences of a top event, or significantly reducing the severity (CCPS & EI 2018).

To rate the effectiveness of a defined barrier, the accompanying text must be used to determine if the barrier was described positively or negatively. This can be achieved through sentiment analysis. An occurrence, as a passage of text may describe the wider situation, and hence sentiment analysis of the entire occurrence may not represent the effectiveness of the extracted barriers. Therefore, the sentiment must be linked to the barrier described within the occurrence. Overall, this is a challenging task as occurrences tend be written in a manner that is very "matter of fact", or features a negative sentiment because something has gone wrong.

To resolve this issue, the paper utilizes the branch of NLP called "Aspect Sentiment Triplet Extraction" (ASTE). ASTE is concerned with extracting aspect terms, opinion terms and sentiment polarity from a given narrative (Peng et al. 2020) (Zhang et al. 2023). The extraction of Aspect Terms can be thought of as acquiring entities such as those defined in sole entity extraction tasks (Baker et al. 2020) (Hughes et al. 2019) (Rybak and Hassall 2021). The Aspect Term can then be categorized, in the case of this paper has a 'barrier'. An Opinion Term associated to the Aspect Term is then identified, this is the term that describes the Aspect Term. Once identified, a sentiment polarity can be derived to establish if the Aspect Term was viewed positively or negatively. Figure 2 demonstrates a basic example of an ASTE system output.

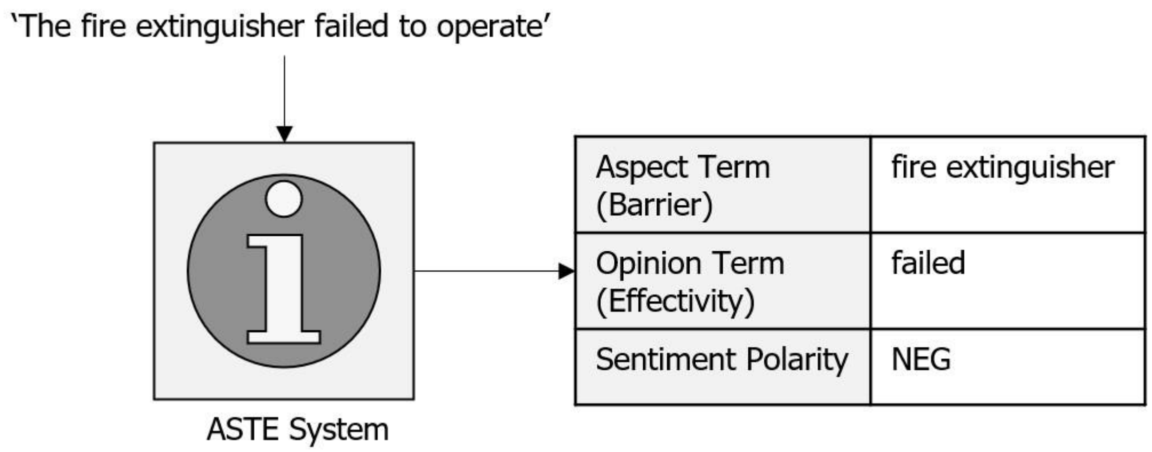

Figure 2 ~ Example ASTE Output from a Simple Occurrence Description

ASTE has attracted much attention within the NLP field in recent years and has utilized several different frameworks and models. Peng *et alia* (2020) first approached the ASTE task with a two-stage pipeline where the first stage extracted aspects, opinion and sentiment (triplet) while the second stage used a classifier to identify aspect-opinion pairs. However, this approach struggled when multiple sentiments were in the text. To overcome this, unified approaches have been proposed extracting triplets in one shot (Wu et al. 2020) (Zhang et al. 2020), or incorporating a position-aware tagging scheme (Xu et al. 2020). A limitation of these works was that they demonstrated reduced performance when the aspect or opinion terms consisted of multiple words—as would be the case with occurrence reports. To overcome this, Xu et al (2021) proposed a model that considered the interactions between spans of aspects and opinions, which forms the core basis of the work described in this paper.

ASTE relies upon a dataset to train a machine learning model—the creation of this is described in the following section. As the MOR data was private and handled under a Non-Disclosure Agreement, only open-source, offline model architectures were considered for this work.

3 Method

An overview of the process followed in this paper is shown in Figure 3 below, while the following sub-sections explain each phase. The raw MOR data was obtained in the form of a large dataframe, where each row represented a single MOR.

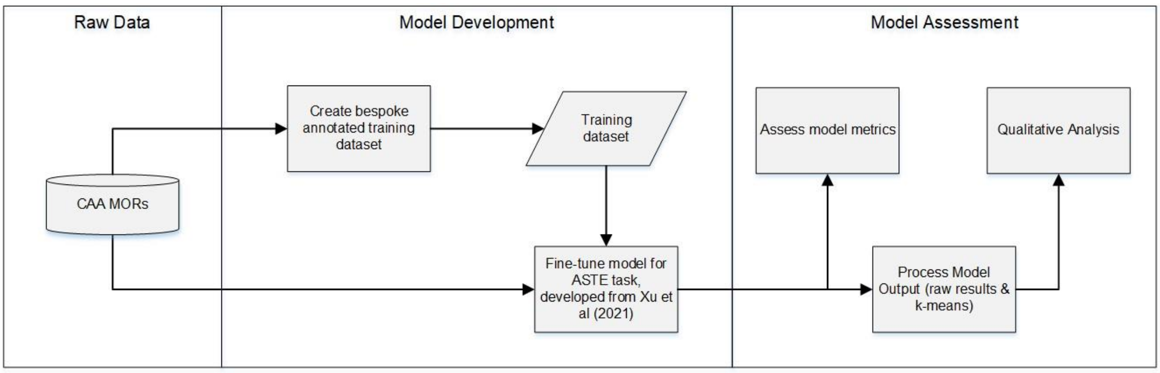

Figure 3 ~ Development Process

3.1 Training Dataset Annotation

At the core of training ASTE models is a specific dataset for the task. Previous ASTE models have been developed from datasets provided in the SemEval series of conferences (Pontiki et al. 2014) (Pontiki et al. 2015) (Pontiki et al. 2016). These datasets feature the annotation of user reviews for electronic devices, laptop computers, and restaurants. As no ASTE dataset exists for assessing occurrences against bowtie barriers, one was developed for this paper.

The bespoke dataset was formatted in the same structure as the SemEval datasets to allow for a wider range of machine learning options while ensuring the SemEval datasets could be easily integrated if a blended dataset were required. The dataset would be used to train the machine learning model to identify barriers, associated effectiveness (opinion terms), and assign sentiment polarity (positive, negative or neutral).

Although software is available to assist with creating datasets, the one developed in this paper was deemed simple enough to be manually created. The occurrences were saved as a column in a dataframe, where two new columns were created; one that separated words and punctuation with white space (e.g. "aircraft," would become "aircraft ," and "26000ft" would be "26000 ft"), and secondly, a column that then annotated the integer position of each word. This allowed word positions and sentiment of barriers to be annotated by the author, who has extensive system safety experience, using Microsoft Excel.

Each occurrence was reviewed to establish if:

1. The text contained a barrier(s), i.e. a hazard control.
2. The text then described the effectivity of barriers.

If both of these were present, the occurrence was annotated and then assigned a sentiment:

- Positive (POS) — The barrier was effective.
- Negative (NEG) — The barrier was not effective.
- Inconclusive/Neutral (NEU) — Not possible to determine barrier effectiveness.

An example of a positive barrier annotation showing the barrier highlighted in green and the effectivity in red is shown below. The resulting annotation would be [([14, 15, 16], [1, 2, 3, 4], 'POS')].

Customer [0] removed [1] one [2] power [3] bank [4] unit [5] from [6] hold [7] baggage [8] after [9] having [10] been [11] shown [12] the [13] pack [14] safely [15] signage [16] . [17]

Any occurrences that featured no barriers were left as a blank list to train the model on such occurrences. Once complete, the dataframe could be processed using the Python *pandas* and *regex* libraries before being saved as a simple *.txt* file matching the format of the SemEval datasets.

The final dataset comprised of 1000 occurrences for training, 150 for development, and 150 for testing. The different types of barriers featured in the training dataset are shown in Figure 4, where the majority of occurrences discussed barriers in a positive sentiment (i.e. the barrier functioned as intended).

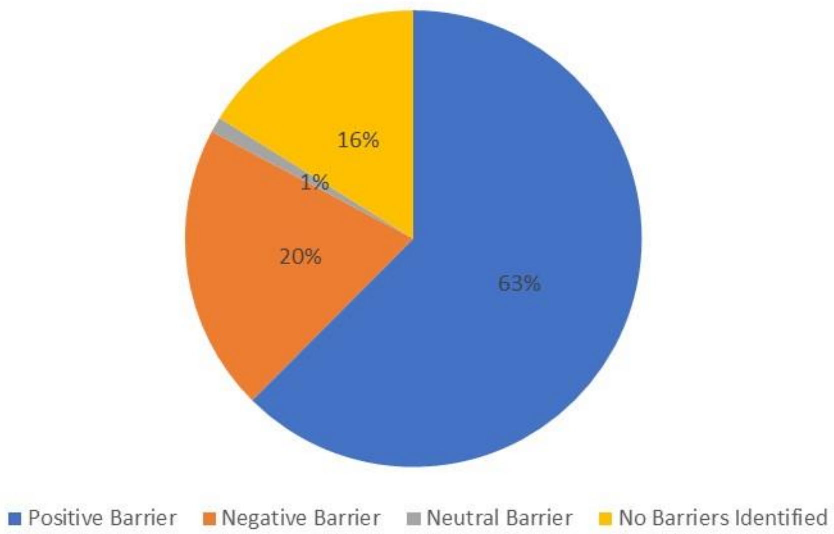

Figure 4 ~ Barrier Sentiments from Training Dataset

3.2 Model Training

The Bidirectional Encoder Representations from Transformers (BERT) model formed the basis of the model developed in this paper, having been pre-trained on the BooksCorpus (800 million words) and English Wikipedia (2,500 million words) (Devlin et al. 2019). The MORs were reviewed and deemed suitable for processing with BERT, due to similarity in language (i.e. only a small percentage of specialist terms or acronyms appeared in the occurrence data).

The BERT model was fine-tuned for the specific ASTE task using the process described by Xu *et alia* (2021). The exceptions being the use of empty lists in the training data to signify occurrences that contained no barriers, and the model parameters modified to handle longer spans of text (up to 16 terms), being that descriptions of effectiveness could be entire sentences.

Model training was completed on a High-Performance Computing suite consisting of 1976 cores with 60 TeraFlops performance.

4 Results

The results of the model were assessed both in terms of model metrics and a qualitative analysis by safety professionals.

4.1 Model Metrics

Being that the dataset and model developed in this paper is the first of its kind, it is not possible to make direct comparisons with other ASTE models. Nor is this appropriate since previous models do not seek to extract barriers or safety-related entities. The metrics from the model are shown in Table 1, it was noted that the metrics were in the similar range as those recorded by Xu *et alia* (2021) for individual SemEval datasets, which provides confidence that model training was successful.

Table 1 ~ Model Metrics

Precision	Recall	F1
0.66	0.41	0.50

The metrics are:

- Precision: Of all the instances that the model identified as Positive, the fraction that were actually Positive;
- Recall: Of all the actual Positive instances, the fraction that the model correctly identified as Positive; and the
- F1 Score is the harmonic mean of Precision and Recall.

The metrics were a result of testing the model against the “test” dataset. The metrics indicate the model appears to be conservative in its predictions, able to correctly identify barriers without identifying non-barriers. However, the recall score indicates barriers may be missed.

4.2 Qualitative Analysis

Assessment of the model metrics represents the initial step to understanding if the model will be useful and warrant further trial. A vital part of the assessment is to gain feedback from safety professionals who would be using the model and reviewing the output. This is more crucial due to the lack of any previous baseline models for comparison.

The model was used to parse a thousand lithium battery related occurrences, producing a dataframe containing the extracted barriers, effectiveness, and associated sentiment. A sample set of these results is shown in Table 2 to demonstrate what is returned by the model.

Table 2 ~ Sample Anonymized Results from Model

	Model Returns		
Occurrence	**Barrier**	**Effectivity**	**Sentiment**
Customer removed from hold luggage one tablet after having been shown the pack safely signage	safely signage	removed from hold luggage one tablet	Positive
During the X ray of passenger's bag, it was found that an e cigarette was inside his hold baggage	X ray	e cigarette was inside his hold baggage	Positive
EMA was untied and without Electric Wheelchair form	Electric Wheelchair form	without	Negative
Smart bag identified during loading. Cptn advised that bag was not allowed to travel as there was no certainty that the lithium battery had been removed from baggage. Cptn overruled tco and allowed baggage to be loaded as baggage owner stated there was no lithium battery inside. This is against [airport] sop for dealing with smart bags	[airport] sop	Cptn over ruled tco	Negative
At the bag drop, ground crew informed the customer about DGR via the safely pack. The customer took out extra batteries he had in his hold bag to put it in his hand bag	ground crew	customer took out extra batteries	Positive

The extracted barriers can be kept within a table, alternatively they may be grouped to form a generalized view on the barriers. Using the Python *sklearn* library, this was achieved through Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to convert the extracted barriers into numerical features. The unsupervised k-means machine learning algorithm was then used to cluster similar barriers together based on these features (Chokor et al. 2016); see Figure 5. Each barrier was then assigned to a cluster before calculating the average sentiment for each cluster. For the purposes of this paper, 10 clusters were selected, however, this can easily be amended should the results require a more precise breakdown.

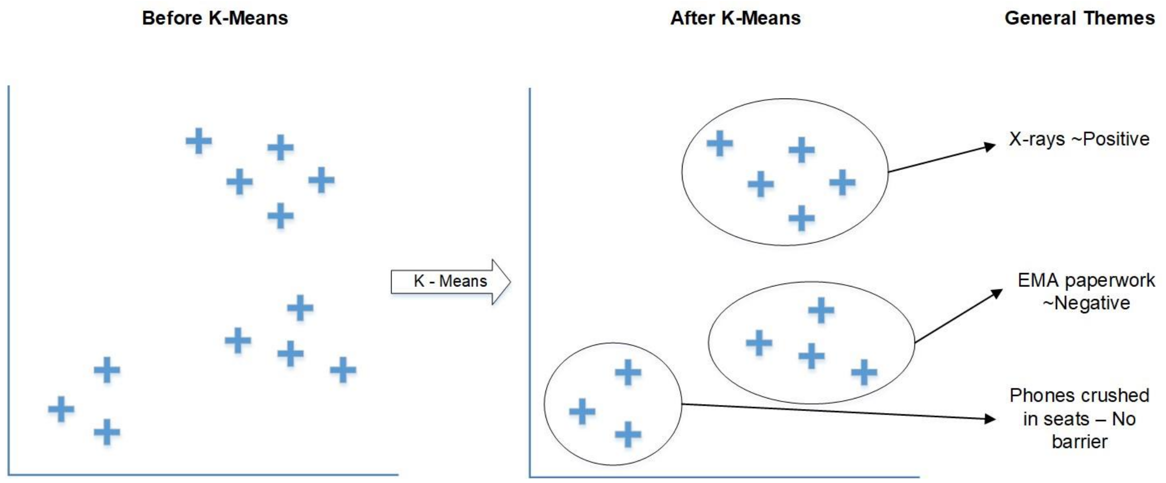

Figure 5 ~ K-means Application to Identified Barriers

As a few examples, the clustered results indicated the following:

- X-ray screening was extracted as a positive barrier, where many occurrences recorded the correct functioning of X-ray screening in detecting lithium batteries. X-ray screening appears as a barrier on the bowtie, hence the method in this paper can be used to evidence the barrier effectivity.
- Notification to Captain (NOTOC) (a document informing the pilot-in-command that dangerous goods are being carried as cargo on board the aircraft) were generally viewed negatively. The main reason being that the NOTOC displayed incorrect information. NOTOCs are not featured as a barrier on the bowtie, although the similar Air Way Bill document is covered.
- A key area where no barriers were identified, nor negative ones returned, was in situations involving the correct stowage of electric wheelchairs (which often contain high-capacity batteries) in the aircraft hold. It was common for the receiving groundcrew to record an occurrence that the wheelchair was found incorrectly stowed post-flight. "Load secured to prevent any movement" is a barrier on the bowtie, hence the model results would drive a negative effectivity for this barrier.
- Ice buckets and metal canisters were returned as positive barriers, because cabin crew use them to immerse electronic devices that are experiencing thermal runaway; however these are not barriers *per-se* on the bowtie.
- No barriers were detected in occurrences that recorded passengers crushing 'phones within their seat mechanisms, suggesting there were limited controls to prevent this hazard from occurring.

The processed output was assessed by five safety professionals via a basic assessment, that requested a score from 1 (strongly disagree) to 5 (strongly agree) against the attributes and statements shown in Table 3 overleaf. Five was deemed the optimal number of users for usability-type testing based upon research by Nielsen *et alia* (1993).

Overall, users liked the consistency of the model output and the insight it provided with the extracted barriers and effectivity. The downsides were the inability to extract all barriers (e.g. some nuisance results could be returned), and the uncertainty that it would reduce workloads. There is still an element of data presentation activity with selecting the number of k-means clusters and reviewing the output — however this is acceptable for a large dataset, and further automation would remove the human element; that is an important aspect for safety related tasks.

Table 3 ~ Qualitative Analysis Results

Attribute	Statement	Average Score
Quality	The output has accurately extracted barriers and effectiveness.	3.4
	The output was consistent.	4.6
	I felt confident using the output.	3.4
Bowtie Application	The output can be easily applied to the existing bowtie.	3.4
	I have gained insight into new barriers that I was previously unaware of.	4.0
	I have gained insight into the effectivity of barriers that I was previously unaware of.	4.2
Workload	Using the output reduces my workload as an analyst.	3.4
	I would not need the support of a technical person to use this output.	3.8
	I anticipate that most safety professionals would learn to use this output very quickly.	4.0

5 Discussion

5.1 Applicability

The output from this work produced a dataframe of extracted barriers and associated effectiveness. This achieved the aim, and allowed safety professionals to understand quickly if barriers are effective and performing as expected—clustering the results allows for a quicker, generalized understanding. The method was suggested to be a useful "assistive model" to help the safety professional, and to supplement a Safety Management System.

One of the main features apparent from creating the dataset, and the model results, was that many extracted barriers were different to those within the carriage of lithium batteries bowtie. The extracted barriers are (mostly) more exact, and pinpoint specific barriers that were encountered by the reporter, while the bowtie resides at a higher level with more generalized barriers. This is not necessarily a bad thing, as barriers are still being identified, however effort is required to compare the model results to the bowtie. This also demonstrates how bowties are an abstract model that can be recorded and drawn in a variety of ways without necessarily being wrong.

A likeliness can be drawn to survivorship bias and the work of Wald (1980), where attention should be paid to the occurrences that feature no barriers—these are areas that seemingly had no defences to hazards and consequences. It would be expected that the Safety Management System should pay close attention to these areas, for example, the discovery of incorrectly stowed electric wheelchairs, or passengers crushing 'phones in seat mechanisms.

It can be difficult for a model to identify a “barrier” being that these are not a single defined entity. A barrier taken from an occurrence may be represented as any part of speech or, more commonly, various parts of speech—especially if including the effectiveness.

The method may work efficiently as part of a barrier management system, or real-time dashboard, where occurrences could be reviewed live and users alerted if given thresholds are breached (e.g. a particular barrier starts featuring sequential negative sentiment).

5.2 Limitations

Creating the training dataset proved to be problematic, whereby it was not always easy for the annotator to determine a barrier. What is perceived as a barrier to one safety professional is not necessarily a barrier to another, hence this created ambiguity. The dataset was also resource intensive to create where, on average, it took 1 person 1 hour to annotate 25 occurrences. This was due to the time taken to read and interpreting the text and possibly checking against the bowtie.

A barrier is sometimes not mentioned at all during an occurrence. For example, some occurrences discussed mobile ‘phones that were crushed within aircraft seats, however there is no “barrier” to tag in this occurrence although there is a genuine risk. Hence, reliance is on investigating such occurrences that mention no barriers.

A number of processed occurrences returned no barriers. However, upon review, barriers were present in the text. Hence, the model is not foolproof and does miss some barriers, which could be problematic, and therefore demands human safety professional oversight. It could be argued that, if the occurrence data is so vast, this may not be an issue because the model processes occurrences much faster than a human, who would still feature an error rate too.

A dependency within this work was that the occurrences supplied by the CAA were already tagged as being lithium battery related, hence a prior classification step had been undertaken prior to model deployment. It is anticipated that if the model in this paper was deployed on a batch of miscellaneous occurrences, then the results could be problematic as the model would require further training data on the new occurrences.

As alluded to within Section 2, it is never possible to claim full validation of a bowtie, and this work demonstrates how different barriers are returned by the model that are not on the set bowtie. It could be possible to add a classification step to assign extracted barriers to those in the bowtie; however this was deemed futile, as bowties typically evolve and change over time. In reality, limited effort is required to review extracted barriers against the bowtie manually.

Ultimately it was not possible to remove analysis by the safety professional. However, given the sheer quantity of occurrence reports the method provides a useful tool for disseminating barrier information from large quantities of occurrences.

5.3 Recommendations and Future Work

An obvious recommendation is to repeat the method on a larger, more diverse, training dataset—i.e. a general set of occurrences, rather than limited to a specific issue such as lithium battery carriage. However, it is the dataset that represents the resource-intensive element of a method, such as ASTE discussed in this paper. Hence, a larger dataset would ultimately be more expensive, using more specialist resource (i.e. safety professionals).

Where this work looked towards bowtie validation, it would equally be useful when constructing a bowtie from scratch. The positive barriers could be built into the bowtie while effort could be assigned to changing negative barriers to positive.

In terms of improving the occurrence data, a set structure for occurrence reports would be highly beneficial if NLP is to be used on a greater scale. Ideally, short-form narratives in a format that indicates the consequence, barriers (and effectiveness), followed by the perceived cause would ensure that only the pertinent information is available while the consistent format would help machine learning models detect and recognize patterns.

Such a structure format of course relies on reporters being aware of bowties, and having a basic knowledge of what constitutes a hazard control/barrier, etc. Consequently, training of reporters could be undertaken to demonstrate what to report and how to report it.

Beyond the focus of one element of a bowtie, future work may wish to consider the generation of bowties from occurrence and incident datasets. It is anticipated that this would require extensive resource to not only extract different unique entities, but also include the ability to link these entities into a sequence.

6 Conclusion

This paper introduces a method and model for extracting bowtie barriers, their effectivity, and sentiment from free-text occurrence reports. The output can be used in the validation, review or creation of bowties, alternatively it may be used to assist a Safety Management System, highlighting areas where barriers are, or are not, effective. The paper focused upon lithium battery related occurrences and barriers; hence the resulting model is limited to these, and represents a capability demonstration for using the technique in wider topics and safety management.

It has not been possible to directly compare this model with previous ASTE works being that they feature different data, and aim to achieve different results. As far as this author is aware, no such prior dataset of barrier effectivity exists. The method was demonstrated to return useful information, allowing the ability to trend and group barriers which would otherwise have required human review—a time consuming and resource intensive task. The results also influenced the bowtie through identifying reported actions that were not captured on the bowtie as barriers.

It is expected that this method could be improved and expanded within the aviation industry or applied to other safety critical industries that feature large quantities of textual data. The model represented a useful "assistive" technique for the Safety Management System; should more automation be required and the model designated 'safety critical', then a larger dataset would be required alongside defined targets to achieve in terms of Precision, Recall and F1 scores.

Data Availability

The datasets presented in this article are not readily available due to Civil Aviation Authority (CAA) UK Regulation 376/2014 that protects data from unauthorised use or disclosure. Requests to access the datasets should be directed to the CAA.

Acknowledgments

J Ricketts thanks the contribution of the Institution of Mechanical Engineers (IMechE) Whitworth Senior Scholarship Award and the UK Civil Aviation Authority in supporting this research.

References

Baker H, Hallowell M. R, and Tixier A. J-P. (2020). *Automatically learning construction injury precursors from text*. Automation in Construction, Volume 118, October 2020, 103145.

CAA. (2023). Occurrence reporting Guidance on mandatory and voluntary occurrence reports (MORs and VORs). United Kingdom Civil Aviation Authority. https://www.caa.co.uk/our-work/make-a-report-or-complaint/report-something/mor/occurrence-reporting. Accessed: 25 August 2025.

CCPS & EI. (2018). *Bow Ties in Risk Management: A Concept Book for Process Safety*. Center for Chemical Process Safety (CCPS) and the Energy Institute (EI). John Wiley & Sons, New Jersey.

Chen Y, Kang Y, Zhao Y, Wang L, Liu J, Li Y, Liang Z, He X, Li X, Tavajohi N, and Li B. (2021). *A review of lithium-ion battery safety concerns: The issues, strategies, and testing standards*. Journal of Energy Chemistry, 59 Science Press, pp. 83–99.

Chokor A, Naganathan H, Chong W. K, and El Asmar M. (2016). *Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning*. International Conference on Sustainable Design, Engineering and Construction. Procedia Engineering, 145 Elsevier B.V. pp. 1588–1593.

Devlin J, Chang M. W, Lee K, and Toutanova K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* In: Burstein J, Doran C, and Solorio T. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1. Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4171–4186.

Hughes P, Shipp D, Figueres-Esteban M, and van Gulijk C. (2018). *From free-text to structured safety management: Introduction of a semi-automated classification method of railway hazard reports to elements on a bow-tie diagram*. Safety Science, Volume 110, Part B, 2018, pp. 11–19.

Hughes P, Robinson R, Figueres-Esteban M, and van Gulijk C. (2019). *Extracting safety information from multi-lingual accident reports using an ontology-based approach*. Safety Science, 118(May) Elsevier, pp. 288–297.

ISO 31010. (2019). *Risk management – Risk assessment techniques*. ISO 31010, 2nd Edition, 2019. International Organization for Standardization, Geneva.

Liu B. (2012). *Sentiment Analysis and Opinion Mining*. (Synthesis Lectures on Human Language Technologies, 5(1)), Morgan & Claypool, San Rafael, CA.

Nazir A, Rao Y, Wu L, and Sun L. (2022). *Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey*. IEEE Transactions on Affective Computing, 13(2) IEEE, pp. 845–863.

Nielsen J, and Landauer T. K. (1993). *Mathematical model of the finding of usability problems*. Proceedings of the Conference on Human Factors in Computing Systems, Amsterdam, pp. 206–213.

Peng H, Xu L, Bing L, Huang F, Lu W, and Si L. (2020). *Knowing what, how and why: A near complete solution for aspect-based sentiment analysis*. AAAI 2020 — 34th AAAI Conference on Artificial Intelligence, pp. 8600–8607.

Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H. Androutsopoulos I, and Manandhar S. (2014). *SemEval-2014 Task 4: Aspect Based Sentiment Analysis*. In: Nakov P, and Zesch T. (eds.) *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland. Association for Computational Linguistics, pp. 27–35.

Pontiki M, Galanis D, Papageorgiou H, Manandhar S, and Androutsopoulos I. (2015). *SemEval-2015 Task 12: Aspect Based Sentiment Analysis*. In: Nakov P, Zesch T, Cer D, and Jurgens D. (eds.) *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).* Denver, Colorado: Association for Computational Linguistics, pp. 486–495.

Pontiki M, Galanis D, Papageorgiou H, Androutsopoulos I, Manandhar S, Al-Smadi M, Al-Ayyoub M, Zhao Y, Qin B, De Clercq O, Hoste V, Apidianaki M, Tannier X, Loukachevitch N, Kotelnikov E, Bel N, Jiménez-Zafra S. M, and Eryiğit G. (2016). *SemEval-2016 Task 5: Aspect Based Sentiment Analysis*. In: Bethard S, Carpuat M, Cer D, Jurgens D, Nakov P, and Zesch T. (eds.) *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* San Diego, California: Association for Computational Linguistics, pp. 19–30.

Prenninger W, and Pretschner A. (2005). *Abstractions for Model-Based Testing*. Electronic Notes in Theoretical Computer Science, Vol. 116, pp. 59–71. Elsevier, Amsterdam.

Reason J. (2000). *Human error: Models and management.* British Medical Journal, BMJ 2000; 320 :768.

Ricketts J, Pelham J, Barry D, and Guo W. (2022). *An NLP framework for extracting causes, consequences, and hazards from occurrence reports to validate a HAZOP study*. IEEE/AIAA 41st Digital Avionics Systems Conference (DASC). Portsmouth, VA, USA: IEEE, pp. 1–8.

Ricketts J, Barry D, Guo, W, and Pelham J. (2023). *A Scoping Literature Review of Natural Language Processing Application to Safety Occurrence Reports*, Safety, 9(2), p. 22

de Ruijter A, and Guldenmund F. (2015). *The bowtie method: A review*. Safety Science, 88 Elsevier Ltd, pp. 211–218.

Rybak N, and Hassall M. (2021). *Deep learning unsupervised text-based detection of anomalies in U.S. Chemical safety and hazard investigation board reports*. International Conference on Electrical, Computer, Communications and Mechatronics Engineering, ICECCME 2021. IEEE, pp. 7–8.

Wald A. (1980). *A Reprint of 'A Method of Estimating Plane Vulnerability Based on Damage of Survivors'*. Defense Technical Information Center, Dayton, OH.

Wu Z, Chengcan Y, Fei Z, Zhifang F, Xinyu D, and Rui X. (2020). *Grid Tagging Scheme for Aspect-oriented Fine-grained Opinion Extraction*. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2576–2585.

Xu L, Li H, Lu W, and Bing L. (2020). *Position-aware tagging for aspect sentiment triplet extraction*. EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, , pp. 2339–2349.

Xu L, Chia Y. K, and Bing L. (2021). *Learning span-level interactions for aspect sentiment triplet extraction.* ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pp. 4755–4766.

Zaratiana U, Tomeh N, Holat P, and Charnois T. (2023). *GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer*. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers), pp. 5364–5376, Mexico City.

Zhang C, Li Q, Song D, and Wang B. (2020). *A Multi-task Learning Framework for Opinion Triplet Extraction*. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 819–828.

Zhang W, Li X, Deng Y, Bing L, and Lam W. (2023). *A Survey on Aspect-Based Sentiment Analysis : Tasks, Methods, and Challenges*. IEEE Transactions on Knowledge and Data Engineering, 35, pp. 1–21.

Zhou W, Zhang S, Gu Y, Chen M, and Poon H. (2023). *UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition*. https://doi.org/10.48550/arXiv.2308.03279