safety

MDPI

Article

Multi-Label Classification of Contributing Causal Factors in Self-Reported Safety Narratives

Saul D. Robinson

Parks College of Engineering, Aviation and Technology, Saint Louis University, Saint Louis, MO 63103, USA; saul.robinson@slu.edu; Tel.: +1-314-977-8252

Received: 13 March 2018; Accepted: 18 July 2018; Published: 20 July 2018

Abstract: Three methods are demonstrated for automated classification of aviation safety narratives within an existing complex taxonomy. Utilizing latent semantic analysis trained against 4497 narratives at the sentence level, primary problem and contributing factor labels were assessed. Results from a sample of 2987 narratives provided a mean unsupervised categorization precision of 0.35% and recall of 0.78% for contributing-factors within the taxonomy. Categorization of the primary problem at the sentence level resulted in a modal accuracy of 0.46%. Overall, the results suggested that the demonstrated approaches were viable in bringing additional tools and insights to safety researchers.

Keywords: latent semantic analysis (LSA); adaptive taxonomy; safety; automatic indexing; machine learning

1. Introduction

Aviation has moved quickly to being one of the safest modes of transportation. In this century-long history, improvements have often resulted from lessons learned as a result of accidents. Where accidents are frequent or directly attributable, this approach has served the industry well. However, great success with this method has resulted in aircraft accidents becoming infrequent and centered on complex human failings. The study of human error that results in accidents has long been an ad hoc assessment of accident and incident data [1].

The challenge of learning from errors has been compounded by both the reduction in accidents and the simultaneous increase in error data. Where accidents may provide clear guidance of what not to do, volumes of data concerning "what could have gone wrong", complicate assessment [2]. The Administrator of the Federal Aviation Administration (FAA) in a 2016 speech highlighted the importance of this challenge to aviation safety. Addressing the World Aviation Training Conference he said, "Simply put, using [big] data to make decisions based on risk is the way of the future" [3].

To date, the use of natural language programming (NLP) and machine learning techniques in the study of safety are in their infancy. Within aviation safety, several authors have shown the usefulness of latent semantic analysis [4] and latent Dirichlet allocation [5] in the reduction and manipulation of narrative data. With differing levels of success, applications have included simple taxonomy classification, document visualization, and the identification of emergent themes [6–8]. Following in this vein, this paper examines the use of latent semantic analysis (LSA) to automatically classify textual narratives from the Aviation Safety Reporting System (ASRS) [9] at the two existing levels of the ASRS taxonomy, using both complete documents and individual sentences from within those documents.

2. Literature Review

Developed by Furnas et al. [10] initially and subsequently popularized by Deerwester et al. [4], latent semantic analysis (LSA) is a mathematical technique for inferring relations between words within bodies of text. As a statistical approach, it makes no assumptions on the construction of the documents beyond the term-frequency counts of the bag-of-words representation. Given vector-space

matrix representation of a corpus, principal values amongst documents and words may be discovered through application of a singular value decomposition (SVD) operation.

The central matrix from the SVD is then truncated by the substitution of the lowest values with zero. This truncated, or reduced space, form of the matrix then provides the inferred relationships between terms used in similar contexts. In this reduced space, term associations are made that are not present in any single body of text. Thus, latent relationships are revealed. In the truncated form, LSA may be used to identify thematic trends (orthonormal vectors of the matrix) or to retrieve information.

Following the reduction of the document to a truncated vector by LSA, a common method to identify and retrieve documents similar to a query, is the unit vector dot product or cosine. This value, between -1 and 1, represents the document similarity within the multi-dimensional space. Where two documents return a cosine value approaching unity, those documents are increasingly identical. Where this comparison is conducted for a set of documents, the document with the highest cosine value is termed the "nearest neighbor".

Robinson, Irwin, Kelly, and Wu [6] analyzed each ASRS narrative as a single object. Both the training corpus, used to develop the LSA space, and the queried document comparisons were treated in this manner. In this work, documents from the query corpus were compared against the training corpus and the label of the nearest neighbor assigned to the query. While moderately effective, this approach was only applicable to the assignment of a single label and ignored the secondary level of the taxonomy.

The observation that an incident narrative forms a story, with a beginning, middle, and end, highlights the limitation of treating them as a single object. Thus, the manipulation of narratives at a reduced level was sought whilst retaining the simplicity of the LSA nearest neighbor approach, such that key elements would not be lost. Due to the absence of paragraph breaks within the ASRS database, the reduction of the narratives was explored at the sentence level. On this basis, our research question was:

- For problems of retrieval within a multi-label taxonomy, can the breakdown of narratives into individual sentences be leveraged to provide new strategies for unsupervised classification?

2.1. Multi-Label Measures

When a narrative report is provided to ASRS, depending on its attributes, it is evaluated by several subject matter experts at NASA. These experts may assign a single label to the *primary problem* field and multiple labels to the *contributing factors* fields.

The ASRS taxonomy provides eighteen available labels (or *causes*), as shown in Table 1, to assign to the *primary problem* and *contributing-factors* fields. Hence, a document may have no relation, with zero labels designated, or include all causes, having seventeen labels allocated. In a combinatorial sense, there are 131,072 possible sub-sets of causal labels available. However, in practice, the usage of all eighteen labels in combination is highly unlikely. Within ASRS, it is common for three labels to be assigned providing 816 combinations or 4896 permutations. For a multi-label problem of this type, there are a limited number of metrics available [11]. Those metrics considered here as most applicable to this type of classification are Hamming loss (H_L) and F_1 score.

Hamming loss, first described by Hamming [12] and defined here in the same manner as Schapire and Singer [13] (see Equation (1)), is the scaled position difference between two binary strings of equal length. Therefore, Hamming loss is bound between zero and unity, where zero indicates no difference between the strings.

$$H_L = \frac{S_i \oplus S_j}{len(i)}; \quad \text{where } i = j \tag{1}$$

The F_1 score, shown by Equation (2), is the harmonic mean of precision and recall. For a given result, the F_1 score has a range of zero to unity, where a score of unity would indicate a perfect

match. When calculating F_1 score, recall (R) is the true positive rate and precision (P) is the positive predictive value.

$$F_1 = 2\frac{P \cdot R}{P + R} \tag{2}$$

Using these measures, the effectiveness of a given classifier—as compared to human-coded labels—within both levels of the ASRS taxonomy may be evaluated.

Table 1. The fractional portion of each label in the training and query datasets. Extrema of each column are shown in bold font.

Labels	Training Fraction		Query Fraction	
	Primary	**Contributing**	**Primary**	**Contributing**
"Ambiguous"	0.123	**0**	0.106	**0**
"Incorrect/Not Installed/Unavailable Part"	0.003	0.011	0.005	0.011
"ATC Equipment/Nav Facility / Buildings"	0.011	0.011	0.007	0.01
"Human Factors"	0.229	0.223	0.265	**0.264**
"Logbook Entry"	**0.000**	0.01	**0.002**	0.01
"Chart Or Publication"	0.013	0.047	0.02	0.04
"Equipment/Tooling"	0.004	0.007	0.004	0.009
"MEL"	0.007	0.013	0.006	0.014
"Airport"	0.012	0.027	0.011	0.034
"Aircraft"	**0.412**	**0.28**	**0.364**	0.26
"Weather"	0.026	0.044	0.025	0.042
"Staffing"	0.001	0.007	**0.001**	0.009
"Environment-Non Weather Related"	0.021	0.044	0.021	0.038
"Company Policy"	0.049	0.07	0.083	0.099
"Manuals"	0.011	0.032	0.006	0.025
"Procedure"	0.073	0.164	0.061	0.12
"Airspace Structure"	0.002	0.011	0.004	0.015
"(No label assigned)"	0.003	0.001	0.009	0.002

2.2. Strategies for Multi-Label Retrieval

The ASRS taxonomy provides eighteen labels for the two categories, *primary problem* and *contributing causes*. Of these categories, the first may contain only one label and the second any combination of the labels. The layers and available combinations of this taxonomy provide a unique problem with no openly available solution strategies from the information retrieval literature.

Since narratives commonly contain multiple sentences, we may assume that each sentence correlates to some degree with the holistic interpretation which precipitated a given set of labels. Following from this assumption, we may utilize the labels of each sentence to select one or multiple as necessary for the taxonomy's categories. In the reduction to one label, we may take those singular labels obtained from each sentences *primary problem* and select only the most frequent as the likely label. In this way, one limiting aspect of treating the narrative as a single object may be overcome. Where multiple labels are to be retrieved, we may use only the singular labels from each individual sentence's *primary problem* as a set. Alternately, the unique labels from each sentences *contributing causes* many be used.

The practical application and calculation of each of these approaches is shown in Appendix A for a single narrative.

3. Method

Self reported safety narratives were taken from the Aviation Safety Reporting System (ASRS) database operated by the National Aeronautics and Space Administration (NASA). Chosen from this database were reports originating from passenger carrying flights operated commercially under Federal Aviation Regulation Part 121, over the period of January 2011 to January 2013. This date range

was selected both to provide a manageable dataset size and to accommodate NASA's server download constraint of 5000 records. This resulted in 4497 reports available to create a primary corpus for LSA analysis. A further sample was taken to generate a query corpus to use for cross-validation. In total, 2987 reports resulted, with the same filter, from the date range of January 2009 to December 2009.

Both corpora were broken down into sentences, using the Natural Language Toolkit *punkt tokenizer* algorithm [14,15]. These sentence sets, which formed the analysis corpora, were then indexed against the document from which they originated. This facilitated the comparison of sentence retrieved to its source narrative's labels. The primary corpus was converted into the bag of words format and stop-words [16] and those words occurring less than 20 times in the corpus were removed. A dictionary of 3663 words remained for document characterization. This process resulted in 63,585 sentences in the training corpus and 43,313 sentences in the query corpus. The individual sentences of the training corpus were used to form an LSA space, truncated to 400 semantic topics (retaining 90.1% of the energy spectrum) following the rationale of Robinson et al. [6], for subsequent document comparison.

This process resulted in a vector representation of each sentence, the document which contained the sentence, and the labels assigned to that document. Subsequently, three strategies were applied to classify the *primary problem* and *contributing factors*. First, it was examined whether the *primary problem* label of a query document could be accurately determined from the nearest neighbor sentence labels of the training corpus. Second, the *contributing factors* of a query document were labeled based on the *primary problem* of nearest neighbor sentences from the training corpus. Finally, the *contributing factors* of a query were labeled based on the *contributing factors* of its nearest neighbors sentences from the training corpus. A demonstration of the three methods is shown in Appendix A for a single narrative.

3.1. Document Primary Problem Evaluation by Sentence Primary Problem

This task required that a single label (*primary problem*) be selected from a list of labels retrieved from each sentence found in the narrative. This was achieved by the selection of the most frequently occurring label in the retrieved list. The process then may be broken down into three discrete steps.

The sentences of the narratives within the query corpus were compared against the training corpus in the LSA space. The *primary problem* of the nearest neighbor sentence from the training corpus was retained. This process was repeated for each sentence within the query document. The result was a list of *primary problem* labels from the nearest neighbor for each sentence in the query document. The most frequently occurring label from this list was then compared against the human-coded *primary problem*.

3.2. Document Contributing Factors Evaluation by Sentence Primary Problem

The *contributing factors* label within the ASRS taxonomy may take on any combination of the eighteen labels found in Table 1. In this approach, the singular labels retrieved from each sentence's *primary problem* was used to populate a list of labels that was then interpreted as the *contributing factors* list. Three individual steps were required.

The sentences of the narratives within the query corpus were once more compared against the training corpus in the LSA space. The *primary problem* of the nearest neighbor sentence from the training corpus was retained. This process was repeated for each sentence within the query document. The result was a list of *primary problems* from the nearest neighbor for each sentence in the query document. This list was then compared against a dictionary of all possible labels and a truncated list of the unique labels returned. This list of unique labels was then compared against the human-coded list of *contributing factors*.

3.3. Document Contributing Factors Evaluation by Sentence Contributing Factors

For the final approach, the list of labels retrieved from each sentence's *contributing factors* was used to populate a list of labels which was assigned as the query's *contributing factors* list. Since repeated instances of a given label are likely, only unique labels were retained. The process may be broken down into the following steps.

The sentences of the narratives within the query corpus were compared against the training corpus in the LSA space. The *contributing factors* of the nearest neighbor sentence from the training corpus was retained. This process was repeated for each sentence within the query document. The result was a list of *contributing factors* from the nearest neighbor for each sentence in the query document. This list was then compared against a dictionary of all possible labels and a truncated list of the unique labels returned. This list of unique labels was then compared against the human-coded list of *contributing factors*.

4. Results

The training corpus was characterized by 4497 narratives containing a minimum and maximum of 1 and 105 sentences, respectively. The sentences had a mean of 14.14, variance of 106.67, skewness of 2.34, and kurtosis of 9.24. The query corpus was characterized by 2987 narratives containing a minimum and maximum of 1 and 102 sentences, respectively. The sentences had a mean of 14.50, variance of 103.37, skewness of 2.09, and kurtosis of 8.04.

The training dataset's use of the *contributing causes* showed a minimum of one and maximum of ten labels. The label demonstrated a mean of 2.26, variance of 1.99, skewness of 1.51, and kurtosis of 2.95. The query dataset's use of the *contributing causes* showed a minimum of one and maximum of nine labels. The label demonstrated a mean of 2.14, variance of 1.59, skewness of 1.34, and kurtosis of 2.05. Of the 306 possible two-label permutations, 89 were observed in the training data and 85 in the query data. Where 4896 three label permutations were possible, 289 were identified for the training data and 226 for the query data.

Table 1 shows the fractional usage of each of the eighteen available labels for the *primary problem* and the *contributing causes* for both the training and query datasets.

Table 2 shows the three classification schemes and their accuracy in terms of Hamming loss and F_1 score and its components. A complete demonstration of each label retrieval process, as applied to an arbitrarily report chosen from the query dataset, is shown in Appendix A.

Table 2. Cause match accuracy for the nearest neighbor found in the training corpus between documents in the LSA space. $\bar{H}_{L_{0.05}}$ indicates the Hamming loss for the lower 5% of cases in the query set.

Matching Strategy	$\bar{H}_L$	$\bar{H}_{L_{0.05}}$	$\bar{H}_{L_{0.95}}$	$\bar{F}_1$	Precision	Recall
Primary	0.229	0.056	0.389	0.343	0.215	0.843
Contributing by primary	0.216	0.056	0.389	0.484	0.351	0.781
Contributing by contributing	0.364	0.111	0.611	0.400	0.255	0.935

5. Discussion

The classification of narrative reports into a taxonomy is a difficult one. When such tasks are given to subject-matter experts, the inter-rater reliabilities are considered acceptable if the label accuracy is above 0.70 [17]. The reliability for a complex taxonomy with many categories, such as *contributing cause* prediction, is then only expected to be worse [18]. It is not surprising then that the accuracies of automated classifiers are also relatively low.

The existing level of precision is not indicative of the importance of taxonomy—or equivalently multi-label problems. Effective classification of textually based data remains an active area of research in both computer science and the study of safety taxonomies. The $\bar{F}_1$ scores here are slightly lower than other studies that combined the use of LSA and nearest neighbor labeling [19,20]. In a study of named entity recognition (identification of person, location, and organization), Guo et al. [19] returned $\bar{F}_1$ scores between 0.50 and 0.70, dependent on the domain analyzed.

When labeling the *contributing cause* with the *primary problem* of the nearest neighbor sentence's originating document provided the most accurate $\bar{F}_1$ score of 0.484. Although this value falls below

the threshold of what is considered acceptable for a taxonomy, errors may be introduced at any stage of the automated process.

The categories of the ASRS taxonomy were first defined at the program's introduction in 1976 [9]. As one of the first self-reported safety programs, the framers of the program did not have the luxury of prior experience and thus by necessity defined the taxonomy at its inception. Subsequently, this classification system has not been substantially changed or evaluated for its effectiveness. Without measure of the systems usability or reliability, it is impossible to determine how much error is introduced at this stage. However, some evidence of mis-classification was found by Robinson et al. [6]. When using LSA to classify the *primary problems* of ASRS reports, Robinson et al. demonstrated that very similar narratives, both by cosine value and subject-matter expert review, were classified differently for a number of reports in the sample.

The dimensionality reduction of latent semantic analysis and following nearest neighbor matching provides another source of error. As with many reduction and machine learning approaches, the justification for a mathematical approach, as applied to a given problem, is one of isomorphism. Where an method is found to be similar to a human process in providing useful answers, it is retained. From the methods outlined within this paper, the linearity of the relationship between terms and topics and their decomposition as orthonormal vectors has interpretable meaning, but does not necessarily reflect a theoretical framework that represents the underlying processes.

The saturation level of each label as a function of sample size may also introduce error. The training corpus used for this study contained only 4497 reports. Other similar classification studies, required upwards of 200 thousand cases to produce higher accuracy rates [19,21]. Here, it is suggested that the size of the training corpus is insufficient in providing examples of each label and their combinations to achieve rates of greater accuracy. However, this limitation could be overcome with additional computing power—scaling by approximately $N^2 \times k^3$ [10], where N is the number of terms plus documents and k the number of dimensions used for LSA. Given that the ASRS database currently contains over 1.4 million reports, the difference is not insignificant [22].

Of the three methods for the evaluation of *primary problem* and *contributing causes*, one further limitation was noted for the process of labeling *contributing cause* on the basis of *primary problem*. Since each sentence of the narrative to be labeled is matched with its nearest neighbor in the training corpus, a single label is retrieved for each sentence. Thus, should a narrative be labeled with more *contributing causes* than there are sentences, the automated process is unable to match the human entry.

5.1. NLP as an Combined Approach to Existing Practice

Beaubien and Baker [23] evaluated ASRS' taxonomy along with several other equivalent aviation reporting systems for their usefulness. Using the criteria of Fleishman et al. [24], Beaubien and Baker identified a number of shortfalls in the ASRS taxonomy. Being the first of its kind, ASRS is able to effectively address three of the four Fleishman et al. criteria—namely standardization of terminology, start-up costs, and the size of the user base. It is in its ability to assist in the solving of applied problems that ASRS has stalled. Beaubien and Baker also expressed concerns regarding the validity of the taxonomy, stating that they "have also been unable to identify any data regarding the accuracy or comprehensiveness of their ratings".

Given the limitations of an aging and fixed system in responding to dynamic and emergent threats, an integrated approach to safety management—utilizing NLP to assist the expert analyst—has become an increasingly important strategy at many airlines and government agencies [7,25,26]. Demonstrated has been the use of NLP by experts to analyze coherence, verify original coding, and assist in coding new reports [7,26].

Using NLP, incoming reports which are easily coded—those which are highly likely to be identified correctly—may be filtered and the most troubling cases presented first to the researcher. It is often these difficult to identify cases that deserve the researcher's principal attention [26]. Thus, with limited resources, researchers may focus their efforts in a more impactful way.

Where prior studies have demonstrated the use of NLP for single-label coding problems [26], the present study has shown several methods for addressing taxonomies containing multiple levels. Through such approaches, it may be practical even with complex taxonomies to reduce human costs and improve coding accuracy. One usage scenario is as follows.

Incoming reports are initially screened by NLP. Those reports with strong neighbors in the training set—found by cosine value—are automatically coded. Those which are less likely to be accurately coded are presented as first priority to the safety analyst along with their assigned codes, similarity scores, and links to the text of the nearest neighbors found. The analyst after evaluating the narrative text may decide to agree with the NLP assessment or reassign the codes. Where the text is separated by sentence and assigned codes, the analyst may choose to further investigate the cause, not only on the basis of their interpretation of the narrative and its originating problem, but also its solution highlighted by the pre-masticated data. A further discussion of this approach for an example narrative is shown in Appendix A.

5.2. Limitations

Although significant efforts have been made to critically asses the reports used for this analysis, the results are preliminary and should be interpreted with caution. Similar studies involving classifiers have used larger training datasets, here only 4497 reports were used. Furthermore, the training and query datasets were sampled from different years. Although it remains unclear if there were substantial changes between 2009 and 2011–2013, it is important to note that there is the potential for sampling error.

6. Conclusions

Narratives often provide a wealth of information to safety analysts that is often condensed through the use of a taxonomy. While taxonomies provide effective summarization of states relevant to safety analysis, they are, in practical terms, immutable. Where risk factors change or classifications need adjustment, existing data must be re-evaluated or ignored. For large databases, such as ASRS, the task of re-labeling narratives into revised taxonomy requires considerable effort.

Computer based approaches to natural language programming provide additional tools with which to summarize and classify narratives. The results of this study build on prior work, suggesting that narratives may be broken down into mathematical objects and interpreted in a manner relevant to safety analysts [6,7,25,27]. The present work suggests that breakdown of the narratives to the sentence level may offer additional assistance in coding reports in multi-level taxonomies. Provided with a larger database—and thus the saturation of each label—it is suggested that the classification accuracy of the approach would improve substantially.

Funding: This research received no external funding.

Acknowledgments: The author would like to thank the unknown reviewers for their constructive comments to prior versions of this paper. In addition, the support of Safety's editorial staff was invaluable in bringing this work to publication.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Demonstration of Processes

The following narrative (identified as ACN 819,963 within the ASRS database) was randomly selected to demonstrate each process utilized for primary and contributing cause prediction. This narrative was human-coded as "human factors" for *primary problem* and "human factors" for *contributing factors*.

> On SLC WEVIC 1 RNAV Departure; the modified route on the PDC clearance was incorrectly understood and programmed as WEVIC direct KSINO versus WEVIC 1 KSINO transition thereby bypassing intermediate waypoints on the departure. Shortly after we passed WEVIC waypoint; Center asked if we were on WEVIC 1 Departure at which time we immediately

detected our error and ATC assigned 20 degrees left to rejoin departure. We queried if there were any traffic conflicts and Center replied there were not. If any questions on PDC modified clearance confirm with Clearance Delivery. Do not allow any rush or distraction to take away from the careful step-by-step verification of ROUTE and LEGS page waypoints compared to clearance.

The preprocessing step identified the following five sentences in the narrative:

1. "On SLC WEVIC 1 RNAV Departure; the modified route on the PDC clearance was incorrectly understood and programmed as WEVIC direct KSINO versus WEVIC 1 KSINO transition thereby bypassing intermediate waypoints on the departure."
2. "Shortly after we passed WEVIC waypoint; Center asked if we were on WEVIC 1 Departure at which time we immediately detected our error and ATC assigned 20 degrees left to rejoin departure."
3. "We queried if there were any traffic conflicts and Center replied there were not."
4. "If any questions on PDC modified clearance confirm with Clearance Delivery."
5. "Do not allow any rush or distraction to take away from the careful step-by-step verification of ROUTE and LEGS page waypoints compared to clearance."

Table A2 shows each sentence from the narrative along with the *primary problem* and *contributing factors* of the nearest neighbor sentences found from the training corpus.

Appendix A.1. Primary Problem by Sentence

The nearest neighbor query resulted in the following list of *primary problem* associated with each sentence: "Human Factors", "Procedure", "Human Factors", "Procedure", and "Aircraft." In Table A1, two of the eighteen available labels are assigned (with one true positive, two false positives, and no false negatives) resulting in values of $F_1 = 0.5$, $P = 0.33$, $R = 1.0$ and $H_L = 0.11$.

Appendix A.2. Contributing Factors by Sentence

The nearest neighbor query resulted in the following list of *primary problems* associated with each sentence: "Human Factors", "Procedure", "Human Factors", "Procedure", and "Aircraft." In Table A1, two of the eighteen available labels are assigned incorrectly (with one true positive, two false positives, and no false negatives) resulting in values of $F_1 = 0.5$, $P = 0.33$, $R = 1.0$ and $H_L = 0.11$.

Appendix A.3. Contributing Factors by Contributing Cause

The nearest neighbor query resulted in the following list of *contributing factors* associated with each sentence: ("Human Factors"), ("Human Factors", "Procedure"), ("Aircraft", "Human Factors"), ("Procedure", "Human Factors", "Chart Or Publication", "Airport"), ("Aircraft". In Table A1 four of the eighteen available labels are assigned incorrectly (with one true positive, four false positives, and no false negatives) resulting in values of $F_1 = 0.33$, $P = 0.2$, $R = 1.0$ and $H_L = 0.22$.

Table A1. Boolean truths for each label, retrieved values, and their differences for the narrative taken from ACN 819,963, coded as "human factors" for *primary problem* and "human factors" for *contributing factors*. Here, S represents the string containing the boolean and the subscripts t, r, c, and p, represent the truth, retrieved, primary problem, and contributing factors, respectively.

Labels	True		Retrieved		Difference		
	$S_{t,p}$	$S_{t,c}$	$S_{r,p}$	$S_{r,c}$	$S_{t,p} \oplus S_{r,p}$	$S_{t,c} \oplus S_{r,p}$	$S_{t,c} \oplus S_{r,c}$
"Ambiguous"	0	0	0	0	0	0	0
"Incorrect/Not Installed/Unavailable Part"	0	0	0	0	0	0	0
"ATC Equipment/Nav Facility/Buildings"	0	0	0	0	0	0	0
"Human Factors"	1	1	1	1	0	0	0
"Logbook Entry"	0	0	0	0	0	0	0
"Chart Or Publication"	0	0	0	1	0	0	0
"Equipment/Tooling"	0	0	0	1	0	0	1
"MEL"	0	0	0	0	0	0	0
"Airport"	0	0	0	1	0	0	1
"Aircraft"	0	0	1	1	1	1	1
"Weather"	0	0	0	0	0	0	0
"Staffing"	0	0	0	0	0	0	0
"Environment-Non Weather Related"	0	0	0	0	0	0	0
"Company Policy"	0	0	0	0	0	0	0
"Manuals"	0	0	0	0	0	0	0
"Procedure"	0	0	1	1	1	1	1
"Airspace Structure"	0	0	0	0	0	0	0
"(No label assigned)"	0	0	0	0	0	0	0

Table A2. Unsupervised sentence-wise primary and *contributing factors* for the nearest neighbor found in the training corpus for the narrative taken from ACN 819,963, which was coded as "human factors" for *primary problem* and "human factors" for *contributing factors*.

Sentence	Nearest Neighbor	
	Primary Problem	Contributing Factors
"On SLC WEVIC 1 RNAV Departure; the modified route on the PDC clearance was incorrectly understood and programmed as WEVIC direct KSINO versus WEVIC 1 KSINO transition thereby bypassing intermediate waypoints on the departure."	HF	HF
"Shortly after we passed WEVIC waypoint; Center asked if we were on WEVIC 1 Departure at which time we immediately detected our error and ATC assigned 20 degrees left to rejoin departure."	PR	HF, PR
"We queried if there were any traffic conflicts and Center replied there were not."	HF	AC, HF
"If any questions on PDC modified clearance confirm with Clearance Delivery."	PR	PR, HF, CP, AP
"Do not allow any rush or distraction to take away from the careful step-by-step verification of ROUTE and LEGS page waypoints compared to clearance."	AC	AC

Human factors (HF), Procedure (PR), Aircraft (AC), Chart or publication (CP), Airport (AP).

Appendix A.4. A Discussion of the Process as Applied to the Example Narrative

The chosen narrative provides a coherent example of the insights that may be gained by this pre-mastication of the data by NLP. ACN 819,963 clearly describes a slip by the flight crew in reading their pre-departure clearance, where the difference between being safe and not, was simply the omission of the numeral one. Undoubtably within ASRS taxonomy, and probably most others, this is unquestionably a "human factors" issue. Given that our demonstrated approach correctly identified the *primary problem* and *contributing factors*, but produced a false positive for "Procedure" and "Aircraft", the question is, is there any insight gained?

Taken at the more granular level of the sentence, the picture changes. Here, we can see that the slip on the crew's part indeed caused a procedural error, which, when noticed, caused them to first verify that there was no conflict with other aircraft. The remaining sentences in the narrative concern the crew members reflections on future practice: first, the clearance confirmation procedure, and, second, the manner in which they program the flight management system.

It is at this level that we can see that the narrative is constructed exactly as was requested by NASA, in that it includes "...what you believe really caused the problem, and what can be done to prevent a recurrence, or correct the situation". Here, we see the loss of information caused by measuring the occurrences of a problem without the solutions provided. Even the most experienced flight crews will continue to err. It can then be seen that to attempt to reduce the *human factors* of this event is not where the most gains to safety are to be made. From the perspective of a company's safety department, the most effective recommendations may be to revise procedures—to require crews to verify clearance changes with ATC—or perhaps recommend accelerating an internal program to install new avionics, which mitigate the issue. (In this particular instance, the avionics system currently being widely adopted is the controller pilot data link (CPDL) which provides direct digital communication between ATC and the flight crew.)

At the sentence level, we can see that classification of both the problem and solution are conjoined, given that this *human factors* issue would be effectively mitigated by *procedure* and *aircraft* equipment changes.

References

1. Wiegmann, D.A.; Shappell, S.A. A Human Error Analysis of Commercial Aviation Accidents Using the Human Factors Analysis and Classification System (HFACS). *Aviat. Space Environ. Med.* **2001**, *72*, 1006–1016. [PubMed]
2. Grabowski, M.; You, Z.; Zhou, Z.; Song, H.; Steward, M.; Steward, B. Human and organizational error data challenges in complex, large-scale systems. *Saf. Sci.* **2009**, *47*, 1185–1194. [CrossRef]
3. Federal Aviation Administration (Ed.) *Safety Revolution*; World Aviation Training Symposium, Department of Transportation: Orlando, FL, USA, 2016.
4. Deerwester, S.C.; Dumais, S.T.; Landauer, T.K.; Furnas, G.W.; Harshman, R.A. Indexing by latent semantic analysis. *JASIS* **1990**, *41*, 391–407. [CrossRef]
5. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
6. Robinson, S.D.; Irwin, W.J.; Kelly, T.K.; Wu, X.O. Application of machine learning to mapping primary causal factors in self reported safety narratives. *Saf. Sci.* **2015**, *75*, 118–129. [CrossRef]
7. Tanguy, L.; Tulechki, N.; Urieli, A.; Hermann, E.; Raynal, C. Natural language processing for aviation safety reports: From classification to interactive analysis. *Comput. Ind.* **2015**, *78*, 80–95. [CrossRef]
8. Robinson, S.D. Visual representation of safety narratives. *Saf. Sci.* **2016**, *88*, 123–128. [CrossRef]
9. Billings, C.; Lauber, J.; Funkhouser, H.; Lyman, E.; Huff, E. *NASA Aviation Safety Reporting System*; Technical Report; National Aeronautics and Space Administration: Washington, DC, USA, 1976.
10. Furnas, G.W.; Deerwester, S.; Dumais, S.T.; Landauer, T.K.; Harshman, R.A.; Streeter, L.A.; Lochbaum, K.E. Information retrieval using a singular value decomposition model of latent semantic structure. In Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Grenoble, France, 13–15 June 1988; pp. 465–480.
11. Tsoumakas, G.; Katakis, I. Multi-label classification: An overview. *Int. J. Data Warehous. Min. (IJDWM)* **2007**, *3*, 1–13. [CrossRef]
12. Hamming, R.W. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **1950**, *29*, 147–160. [CrossRef]
13. Schapire, R.E.; Singer, Y. Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.* **1999**, *37*, 297–336. [CrossRef]
14. Bird, S. NLTK: The natural language toolkit. In Proceedings of the COLING/ACL on Interactive Presentation Sessions, Sydney, Australia, 17–18 July 2006; Association for Computational Linguistics: Uppsala, Sweden, 2006; pp. 69–72.

15. Garrette, D.; Klein, E. An extensible toolkit for computational semantics. In Proceedings of the Eighth International Conference on Computational Semantics, Tilburg, The Netherlands, 7–9 January 2009; Association for Computational Linguistics: Uppsala, Sweden, 2009; pp. 116–127.
16. Wilbur, W.J.; Sirotkin, K. The automatic identification of stop words. *J. Inf. Sci.* **1992**, *18*, 45–55. [CrossRef]
17. Wallace, B.; Ross, A. *Beyond Human Error: Taxonomies and Safety Science*; CRC Press: Boca Raton, FL, USA, 2004.
18. Taib, I.A.; McIntosh, A.S.; Caponecchia, C.; Baysari, M.T. Comparing the usability and reliability of a generic and a domain-specific medical error taxonomy. *Saf. Sci.* **2012**, *50*, 1801–1805. [CrossRef]
19. Guo, H.; Zhu, H.; Guo, Z.; Zhang, X.; Wu, X.; Su, Z. Domain adaptation with latent semantic association for named entity recognition. In Proceedings of the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Boulder, CO, USA, 1–3 June 2009; Association for Computational Linguistics: Uppsala, Sweden, 2009; pp. 281–289.
20. Liu, X.; Zhang, S.; Wei, F.; Zhou, M. Recognizing named entities in tweets. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; Association for Computational Linguistics: Uppsala, Sweden, 2011; pp. 359–367.
21. Dredze, M.; Wallach, H.M.; Puller, D.; Pereira, F. Generating summary keywords for emails using topics. In Proceedings of the 13th International Conference on Intelligent User Interfaces, Perugia, Italy, 30 June–3 July 2008; pp. 199–206.
22. National Aeronautics and Space Administration. *ASRS Program Briefing*; Technical Report; National Aeronautics and Space Administration: Washington, DC, USA, 2016.
23. Beaubien, J.M.; Baker, D.P. A review of selected aviation human factors taxonomies, accident/incident reporting systems and data collection tools. *Int. J. Appl. Aviat. Stud.* **2002**, *2*, 11–36.
24. Fleishman, E.A.; Quaintance, M.K.; Broedling, L.A. *Taxonomies of Human Performance: The Description of Human Tasks*; Academic Press: Cambridge, MA, USA, 1984.
25. Halford, C.; Harper, M. ASIAS: Aviation Safety Information Analysis and sharing. In Proceedings of the IEEE/AIAA 27th Digital Avionics Systems Conference, Saint Paul, MN, USA, 26–30 October 2008.
26. Pimm, C.; Raynal, C.; Tulechki, N.; Hermann, E.; Caudy, G.; Tanguy, L. Natural Language Processing (NLP) tools for the analysis of incident and accident reports. In Proceedings of the International Conference on Human-Computer Interaction in Aerospace (HCI-Aero), Brussels, Belgium, 12–14 September 2012.
27. Agovic, A.; Shan, H.; Banerjee, A. Analyzing Aviation Safety Reports: From Topic Modeling to Scalable Multi-Label Classification. In Proceedings of the 2010 Conference on Intelligent Data Understanding, Mountain View, CA, USA, 5–6 October 2010; pp. 83–97.