safety

MDPI

Article

Classifying Aviation Safety Reports: Using Supervised Natural Language Processing (NLP) in an Applied Context

Michael D. New * and Ryan J. Wallace

Aerospace Engineering Department, Daytona College of Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, USA; wallacr3@erau.edu
* Correspondence: newm1@erau.edu

Abstract: This paper presents a practical approach to classifying aviation safety reports in an operational context. The goals of the research are as follows: (a) successfully demonstrate a replicable, practical methodology leveraging Natural Language Processing (NLP) to classify aviation safety report narratives; (b) determine the number of reports (per class) required to train the NLP model to achieve an F_1 performance score greater than 0.90 consistently; and, (c) demonstrate the model could be implemented locally, within the confines of a typical corporate infrastructure (i.e., behind the firewall) to allay information security concerns. The authors purposefully sampled 425 safety reports from 2019 to 2021 from a university flight training program. The authors varied the number of reports used to train an NLP model to classify narrative safety reports into three separate event categories. The NLP model's performance was evaluated both with and without distractor data, running 30 iterations at each training level. NLP model success was measured using a confusion matrix and calculating Macro Average F_1-Scores. Parametric testing was conducted on macro average F_1 score performance using an ANOVA and post hoc Levene statistic. We determined that 60 training samples were required to consistently achieve a macro average F_1-Score above the established 0.90 performance threshold. In future studies, we intend to expand this line of research to include multi-tiered analysis to support classification within a safety taxonomy, enabling improved root cause analysis.

Keywords: aviation safety reports; multi-class; natural language processing (NLP); bidirectional encoder representations from transformers (BERT)

Academic Editor: Raphael Grzebieta

Received: 24 October 2024
Revised: 7 January 2025
Accepted: 14 January 2025
Published: 16 January 2025

Citation: New, M.D.; Wallace, R.J. Classifying Aviation Safety Reports: Using Supervised Natural Language Processing (NLP) in an Applied Context. *Safety* **2025**, *11*, 7. https://doi.org/10.3390/safety11010007

1. Introduction

Thousands of times each year, airline employees file internal safety reports to alert management of incidents and hazards encountered during aviation operations. Most commonly, these reports are submitted via the organization's electronic internal safety reporting system, which generally includes a database for securing, tracking, and managing these submissions. Reports are usually written using "freeform text describing the incident, along with a small set of metadata (mostly concerned with the time, the location, and the equipment involved)" ([1], p. 81). The report is then queued for initial triage by an aviation safety analyst, who reads the report narrative to evaluate the significance of the content and determine how to share the information within the organization.

Several unique challenges complicate the process of analyzing and classifying safety reports. For example, safety analysts may lack direct experience or context in the reported topic areas. Just as analysts develop the ability to recognize genuinely serious events, professional advancement, job turnover, and other factors can rob aviation safety departments of their skillset.

Their format and makeup further challenge the interpretation of safety reports. Unstructured text [2,3], varied writing styles, reporter perspective, and writing spontaneity can make interpreting safety reports difficult (Tanguy et al., 2015). "These texts are written in plain language, show a wide range of linguistic variation (telegraphic style overcrowded by acronyms or standard prose) and exist in different languages, even for a single-company country" ([1], p. 80).

Yet another confounding factor is that safety report content is often overly focused on non-relevant material—content that does little to reveal incident causes or effects [4]. Additionally, analyst perspective and bias in interpreting data may inject ambiguity into the evaluative process [5]. These characteristics, coupled with nuance, emotional language, and other factors, can easily lead to reports being misinterpreted or salient safety details being essentially buried among less relevant content within the text.

Moreover, existing processes for analyzing safety reports are both cumbersome and labor-intensive. According to Ahadh [6], manual analysis "is unattractive since it is expensive, time-consuming, and error-prone" (p. 457). Complicating this problem is that analysts are often responsible for processing large numbers of reports [1].

The paper is organized to include a summary literature review, highlighting important advancements in NLP and ML for narrative analysis, followed by an overview of the current problem, and research questions. The methodology section overviews the current study design considerations, framework, instrumentation, data preparation, and processing. In the results and discussion section, the authors present the model output and assess model performance metrics. We conclude the paper by answering the posed research questions and presenting recommendations and plans to expand this line of research.

1.1. Literature Review

Several prior attempts have been made to leverage machine learning strategies to classify safety-based narrative data, with varying degrees of success. Early classification models treated narrative data as a bag of words—independent occurrences of words within the text, without regard to grammar, word order, structure, or context. The resulting output is a simple string of frequency counts for various features—word selections—found within each narrative [6]. While this approach was among the first successful strategies to transform qualitative data for quantitative analysis, its utility was limited and wrought with significant limitations for interpreting context. The approach used in the Bag of Words analysis would be improved to segregate meaningful words from less meaningful ones. Dubbed Term Frequency-Inverse Document Frequency (TF-IDF), this methodology applied a statistical weighting technique to identify unique terms across an analyzed narrative or corpus. The TF-IDF method yields a word score based on the rarity of the term, essentially representing the importance of the word relative to all narratives in the corpus [7,8]. Despite its improvements over the Bag of Words method, TF-IDF has several limitations in appropriately interpreting narrative context. Narrative context interpretation would improve with the advent of Support Vector Machines (SVMs). SVMs enable further improvement to narrative analysis by providing a means for simple classification by establishing a narrative hyperplane—a mathematical representation of words that enables rudimentary binary classification prediction. While SVMs are robust to handling high-dimensional data—narratives with a large number of features or variables—the approach still lacks the ability to interpret semantic meaning, based on narrative context or relationships [6,9].

Current NLP models apply a neural network, deep learning design, which uses a multi-layered approach modeling functionality similar to that of the human brain. A neural network processes input data through a multi-layered system, which applies transformations, weights, bias adjustments, and activation functions to solve complex narrative issues

such as sophisticated pattern recognition and non-linearity. The application of neural networks to NLP analysis improves semantic interpretation of narratives [10], and demonstrates superior performance over other classification techniques when applied to textual datasets [9]. The improvement in accuracy and performance of neural network models over previous narrative analysis methods make them particularly popular for these tasks [11].

1.2. Problem Statement

The high number of analyst tasks, coupled with limited contextual knowledge, creates conditions that allow events of interest—incidents that could represent a serious risk to organizational safety—to linger unseen or unrecognized in the safety reporting database. Moreover, reports are usually processed in the order received, potentially delaying the review and response to events with significant consequences.

1.3. Purpose Statement

This research aimed to create a practical assessment tool using natural language processing (NLP) for automated processing and classification of narrative-based safety reports to augment safety analyst triage for events of interest.

1.4. Significance of the Study

Applying NLP techniques to the analysis of safety reports does not substantively change the overall processing and disposition of reports. Instead, it streamlines their management [1], accuracy, and triage. According to Paraskevopoulos et al. [4], "combining machine learning with natural language processing can automate their [safety report] classification and help safety managers . . . to quickly understand underlying conditions and factors and gain insights for proper assessment regarding safety measures" (p. 3).

Rickets et al. [11] assert, "Recent advances in deep learning models such as Bidirectional Transformers for Language Understanding are now achieving a high accuracy while eliminating the need to substantially pre-process text" (p. 1). Tanguy et al. [1] propose that NLP processes can also aid analysts by identifying common or routine themes, thereby preserving analyst expertise and effort for more hazardous incidents. Idyllically, achieving high-precision NLP of safety reports could minimize or eliminate human verification, reducing reliance on limited human knowledge and manual review [1,11].

Specifically, integrating NLP into the safety reporting handling process is intended to improve the rapid identification of events of interest, improve the accuracy and consistency of initial classification over existing methods, and reduce the reliance on human analysts (without specific subject matter expertise) in the evaluation process.

1.5. Research Questions

The authors sought to answer the following research questions:

- Can NLP techniques identify events of interest by analyzing aviation safety report narratives?
- How many training samples are needed for the NLP model to consistently achieve a Macro Average F_1-Score of 0.9 on a multi-class classification task?

2. Methodology

2.1. Design Requirements

For this study, the authors identified several requirements that drove the design and implementation of the resulting NLP model. To be effective, the use of NLP for the analysis of safety reports must perform the following:

- Minimize the effort expended by high-value subject matter experts (SMEs) to identify examples of "events of interest" from a large set of safety report narratives.
- Achieve an average F_1-Score greater than 0.9 for identified categories, even when distractors are included. Note: The authors operationally define distractors as events with similar characteristics (such as writing style, lexicon, and length) as the events of interest but are determined by SMEs not to meet the criteria for inclusion in any identified categories.
- Meet equipment and processing requirements so the NLP classification process can be performed on a local operating system for information security reasons.

2.2. NLP Using Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) is the latest iteration of the NLP model design that applies a neural network, deep learning methodology. BERT is a relational language model comprising a transformer, incorporating an encoder designed to interpret and process contextual text and a decoder that performs a task, such as classification. "BERT makes use of the transformer, an attention mechanism that learns the contextual relations between words (or sub-words) in a text" ([12], p. 1). Additionally, BERT is advantaged by evaluating word context in both left and right directionality [13,14], enabling improved contextual accuracy over unidirectional NLP models and performing well compared to other NLP models [9,15].

In its default state, BERT comes pre-trained in essential language functions. However, equipping BERT to classify narratives requires fine-tuning to refine the model further to accomplish specific tasks, such as classification, sentiment analysis, summarization, translation, or related activities.

NLP classifiers, such as BERT, output a proportional distribution that rates the classification fit of each narrative within the classification categories. For many NLP models, classifications are recommended based on the category that receives the highest probability score, regardless of the extent of the difference between the other category scores. This means that a narrative that receives a slightly higher score in one category over another will be classified in that category. NLP classifiers can be configured for binary factor categorization; multi-class, multi-factor classification; or, multi-class, mutually exclusive classification.

2.3. Methodological Framework

This research employed an applied research methodology using a quantitative approach. The authors sampled narrative safety report data from a university flight training program. Three unrelated events of interest were selected for categorization. Subject matter experts (SMEs) evaluated sample report narrative data for applicable categorization. Ensuring the SMEs are properly calibrate is a critical step required when constructing a training dataset [6,16]. To ensure this requirement was met, the research team adapted techniques developed by Holt et al. [17]. Specifically, the SME performance was monitored against a referent classification standard through the use of initial training, calibration sessions, and random checks by the investigator.

The research team varied the number of SME-coded training reports to build the NLP model from 20 to 60 reports in 10-report increments (Independent Variable). Model testing was performed on a randomly sampled subset of data to evaluate model performance. The model performance was tested with and without distractor data—irrelevant, misleading, or confounding data. Following each run of the model testing, an F1-Score was calculated and recorded (Dependent Variable). The authors performed statistical testing on the resulting F1-Scores to identify and assess significant differences in model performance related to the Independent Variable (i.e., the number of reports used to train the model).

2.4. Instrumentation

The authors used BERT to build a classification model to classify safety reports into mutually exclusive, pre-defined categories. The NLP model was developed and tested using a single Razer Blade 18, a high-end gaming laptop with 32 GB of RAM, and an internally installed NVIDIA GeForce RTX 4090 graphics processing unit (GPU, NVIDIA, Santa Clara, CA, USA) to run the BERT transformer.

2.5. Data Sample

The sample pool comprised 425 safety reports purposively sampled over three years, from 2019 to 2021. Subject matter experts reviewed and coded the report narratives until at least 80 cases were identified for each of the three possible selected event categories.

2.6. Data Preprocessing

This model relies on narrative text to properly categorize safety reports. Short report narratives yield a much higher variability in categorization probability, so the authors filtered out reports with less than 25 words in the narrative.

Data must be normalized before analyzing narrative data with NLP processes, removing stop words, text casing, and consolidating word forms or derivatives into a singular form [18]. During this preprocessing stage, punctuation was removed, except for periods. Adadh [2] highlights the importance of developing a field dictionary to aid the NLP model in understanding report corpus or contextual keywords. The research team reviewed sample reports to identify critical, contextual, and commonly used acronyms and abbreviations. A dictionary or lexicon lookup was created to replace common acronyms and abbreviations. Eighty-seven terms were integrated to improve the context recognition of the model. The lexicon included terms that defined crew responsibilities, aircraft positional or maneuver information, airplane and aviation infrastructure systems, and local aerodrome references. Final preprocessing converted all text casing to lowercase. An example of the normalization process adjustments to the safety report narratives is provided in Table 1. (Note: whereas this example contains items similar to the narratives used in the study, it is offered for demonstration purposes only and is not based on an actual safety report, as these are considered confidential information by the institution providing the raw data.)

2.7. Class Definition

The authors established three events of interest referred to throughout this paper as either categories or classes. Narratives related to these classes (and selected distractors) were limited to the takeoff and landing phases of flight:

- Event_0 = Skill-related event;
- Event_1 = Airspace-related event;
- Event_2 = Mechanical-related event.

Each class contained fundamentally different environmental and behavioral components, making each distinctive to the NLP classifier. These classes were also selected due to the preponderance of available safety report data, ensuring the adequacy of events of interest data.

2.8. SME Classification

Subject matter experts (SMEs) manually categorize the safety report narratives. For training the model, codes representing SME categorizations are linked to the raw narrative records, and the resulting Training Dataset serves as the input used by the NLP script to build the classification model. The resulting NLP model enables new safety reports (i.e., not previously categorized by SMEs) to be ingested and yield a probability distribution

indicating the model's recommendation for categorization. Properly configured, this process provides automated flagging of various events of interest identified by the model.

Table 1. Example of the Data Normalization Process.

Original Narrative	Preprocessing	Replaced Abbreviations	Model
I was the PIC and PF on a flight from ABC to XYZ. Shortly after tkof we heard a loud thump sound. I immediately checked our eng instruments and noticed that the RPM gauge was in the red. The PM communicated with ATC, letting them know we were declaring an emergency and needed to return to the dep airport. We performed the chklist and ATC then told us to turn right and climb to 3000 MSL. While in the turn (at about 2600′), we received a traffic alert. We did not see any traffic and continued our climb to the assigned altitude. ATC then provided vectors for the ILS RWY 17. They clred us direct to the IAF and told us to contact the Tower. We contacted TWR and the lndg was normal.	I was the PIC and PF on a flight from ABC to XYZ. Shortly after tkof we heard a loud thump sound. I immediately checked our eng instruments and noticed that the RPM gauge was in the red. The PM communicated with ATC letting them know we were declaring an emergency and needed to return to the dep airport. We performed the chklist and ATC then told us to turn right and climb to 3000 MSL. While in the turn at about 2600 we received a traffic alert. We did not see any traffic and continued our climb to the assigned altitude. ATC then provided vectors for the ILS RWY 17. They clred us direct to the IAF and told us to contact the tower. We contacted TWR and the lndg was normal.	I was the captain and pilot flying on a flight from ABC to XYZ. Shortly after takeoff we heard a loud thump sound. I immediately checked our engines instruments and noticed that the RPM gauge was in the red. The pilot monitoring communicated with ATC letting them know we were declaring an emergency and needed to return to the departure airport. We performed the checklist and ATC then told us to turn right and climb to 3000 MSL. while in the turn at about 2600 we received a traffic alert. We did not see any traffic and continued our climb to the assigned altitude. ATC then provided vectors for the instrument landing system runway 17. They cleared us direct to the IAF and told us to contact the tower. We contacted tower and the landing was normal.	I was the captain and pilot flying on a flight from ABC to XYZ. Shortly after takeoff we heard a loud thump sound. I immediately checked our engines instruments and noticed that the RPM gauge was in the red. The pilot monitoring communicated with ATC letting them know we were declaring an emergency and needed to return to the departure airport. We performed the checklist and ATC then told us to turn right and climb to 3000 MSL. while in the turn at about 2600 we received a traffic alert. We did not see any traffic and continued our climb to the assigned altitude. ATC then provided vectors for the instrument landing system runway 17. They cleared us direct to the IAF and told us to contact the tower. we contacted tower and the landing was normal.

2.9. Multiple-Runs by the Level of the Independent Variable

For each level of the Independent Variable (i.e., number of records used to train the model—20, 30, 40, 50, and 60), the model was run 30 times. Following each training run, 20 reports from each category combined in a testing dataset were presented to the resulting model. Reports used for model testing differ from those used for model training as they were randomly selected from the remaining pool of coded reports after removing the testing narratives.

2.10. Distractors

The design philosophy of this project includes the ability to parse out examples of specific events/hazards for immediate action that may be buried within a large set of narrative data. These narratives were pulled from the same report corpus and evaluated by the SMEs. Although nearly all reports used similar phraseology and some events were similar to the event categories of interest, the SMEs determined they were not members of the targeted classes.

2.11. NLP Model Success Measures

To analyze the performance of the model, the research team used an assessment instrument known as an F-measure, reported as an F_1-Score. The F_1-Score is tabulated by comprehensively comparing true positives, false positives, and false negatives. True Positives (TP) are the number of samples correctly predicted as positive. False Positives (FP) are the number of samples wrongly predicted as positive. False Negatives (FN) are the number of samples wrongly predicted as negative.

The F_1-Score is an algorithm that amalgamates classification model performance using the metrics of precision and recall (Leung, 2022). Precision measures the level of accuracy of the predicted positives. Precision is calculated by determining the proportion of true positives (TP) relative to the number of model-predicted positives (TP + FP). Recall measures the proportion of predicted positives to the total number of available true positives. Recall is calculated by determining the proportion of true positives (TP) relative to the number of true positives and false negatives (TP + FN). See Figure 1.

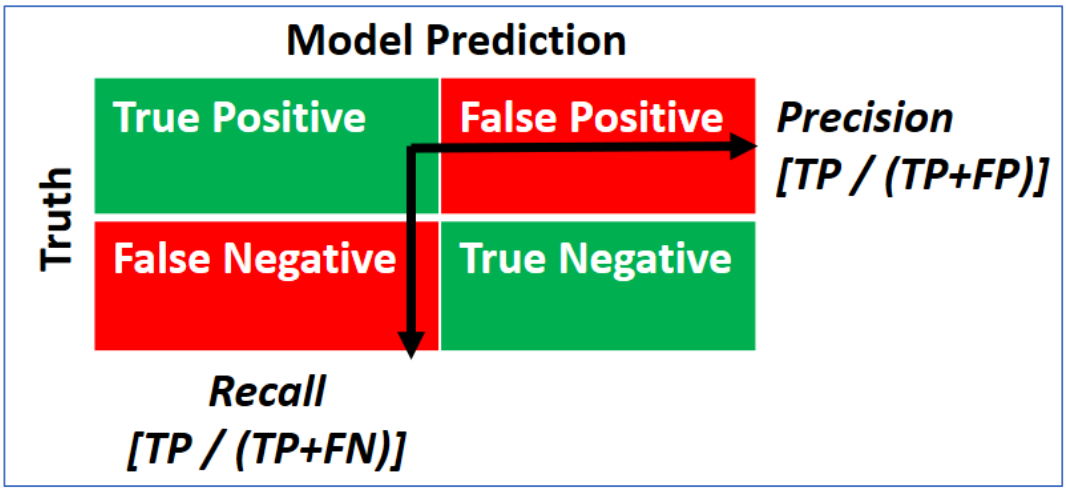


Figure 1. Relationship Between Precision and Recall for Multi-Classification NLP Model Performance. Note: Adapted from Bex [19].

This yields the following F_1-Score equation:

$$F_1\text{-Score} = TP/(TP + \tfrac{1}{2}(FP + FN))$$

F_1-Scores are tabulated for each classification category, with overall model performance reported using a Macro Average, an unweighted algorithmic mean of all category F_1-Scores [20]. While there are several variants for tabulating overall model performance using F_1-Scores, the research team elected Macro Averaging since the sample contained a balanced dataset [20].

2.12. Assumptions and Limitations

The use of NLP classifiers relies on several assumptions and limitations, notably:

- As identified by Yang and Huang [18], aviation is an international industry; therefore, NLP safety analysis methods must be able to support multiple languages and dialects. For this research, reports were limited to those written in English. While English remains the international standard for aviation communications, the authors recognize the need for NLP analysis to be adaptable to varied language, culture, and other considerations. However, these factors were not analyzed during this study.

- Leveraging a supervised training model for NLP classification is not adaptable to changing safety conditions and is not applicable to identifying previously unknown emerging threats [1].

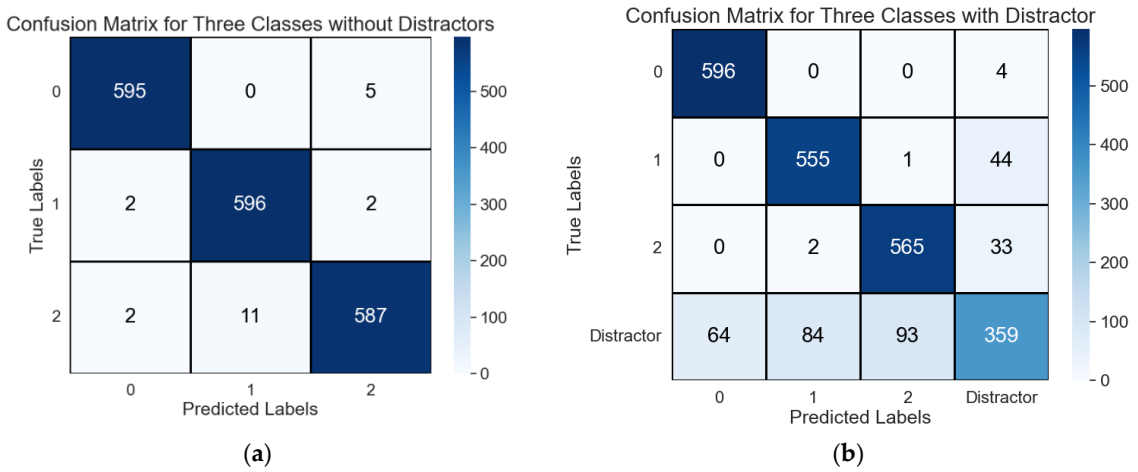
2.13. Institutional Review Board and Participant Protection

This research project utilized data collected from a secondary source. The data are not publicly available, and the provider was not involved in the research project. While the provider can link the data back to living individuals, the data furnished to the authors was de-identified. Based on these factors, it was determined by the Institutional Review Board (IRB) that the research falls under 14 CFR 46 (4), Exempt Secondary Research, for which consent is not required and no IRB review is needed. The original dataset was not released because of the sensitive nature of individual safety reports and organizational safety data used in this project.

3. Results and Discussion

3.1. Confusion Matrices

To explore the general performance of the model and gain insight related to the influence of the distractors, the results from all runs (by the level of the Independent Variable) were combined, and the following confusion matrices were tabulated: (a) model performance without distractors and (b) model performance with distractors. These matrices provide a visual representation of model performance. A sample of results from the 50 training sample levels is provided in Figure 2.



(**a**) (**b**)

Figure 2. Example NLP Model Confusion Matrices at 50 Training Sample Level. (**a**) model performance without distractors and (**b**) model performance with distractors.

3.2. F_1-Score

The authors calculated the Macro F_1-Score after each model run. As mentioned previously, the model was run 30 times at each level of the Independent Variable (IV). The line plot for these results is presented in Figure 3.

When distractor narratives were added that did not correspond to the event categories, the NLP model performance decreased. Adding distractors increases ambiguity, requiring further NLP model refinement, including the addition of more training samples to achieve the desired 0.9 performance level. Initial visual inspection suggests that sustained NLP model performance above the 0.9 threshold may be achieved when using 50 training samples.

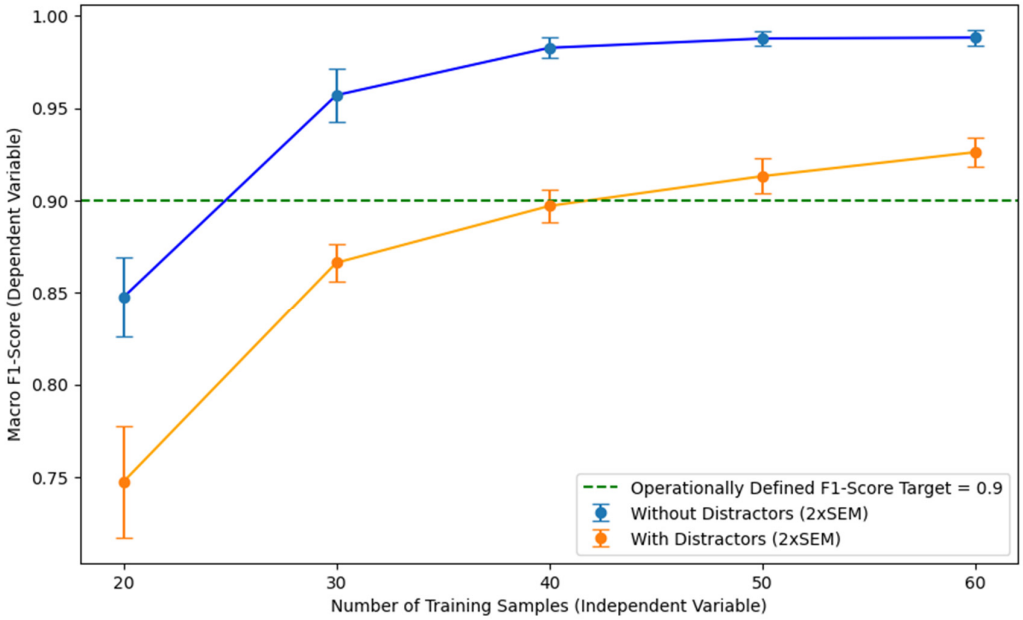


Figure 3. Macro F_1-Scores by Number of Samples Used to Train the Classification Model.

3.3. Statistical Testing of Group Differences Using the Macro-Average F_1-Score

This study's targeted application is to incorporate NLP during the initial filtering stage (i.e., "triage") of report processing. Therefore, the following analysis focused on testing whether significant differences exist between the IV levels as the performance of the model approaches (or exceeds) the operationally defined target of 0.9 F_1-Score when classifying reports in the presence of distractors. See Figure 4 for a graphical depiction of the analysis.

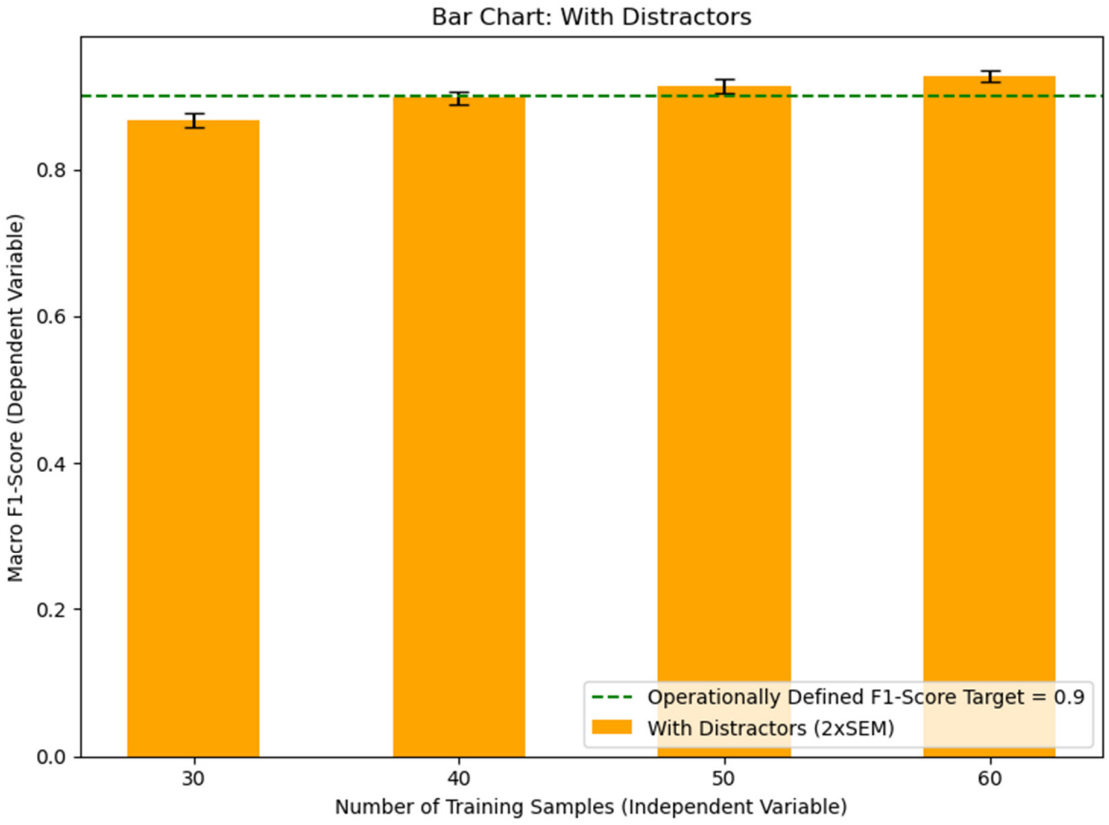


Figure 4. NLP Model Performance Measured by Macro Average F_1-Scores by Training Sample Level.

The prerequisites for parametric statistical testing were not met when including the 20 sample results. Based on the relatively low Macro Average F_1-Score performance of the 20 sample level (as depicted in Figure 3), the authors decided to remove the group from the statistical analysis. When the 20 sample level was removed, the data met criteria for homogeneity and normality.

Statistical testing was performed using an ANOVA, and the Levene test was used to ensure the prerequisite homogeneity of variance assumption was met, F (3, 116) = 1.476, p = 0.225.

The ANOVA test indicated a statistically significant difference in Macro Average F_1-Scores across the four levels of the IV evaluated F (3116) = 32.465, $p < 0.001$. To determine pairwise statistical significance, post hoc comparisons were conducted using Tukey's HSD test and evaluated against an alpha level of 0.05. The results indicated the following (Table 2):

Table 2. Post Hoc Tukey HSD Testing Results.

Group	Significance
30–40	$p < 0.001$ (significant)
30–50	$p < 0.001$ (significant)
30–60	$p < 0.001$ (significant)
40–50	$p = 0.063$ (not significant)
40–60	$p < 0.001$ (significant)
50–60	$p = 0.183$ (not significant)

These findings suggest that 60 training samples are needed to consistently achieve a Macro Average F_1-Score greater than the operationally defined goal of a 0.9 performance threshold.

4. Conclusions and Recommendations

Can NLP techniques identify events of interest by analyzing aviation safety report narratives?

This project demonstrates the capability to use NLP models as a tool for safety analysts to classify and triage high-priority, free-text aviation safety reports for known hazards. NLP modeling bridges the knowledge gap between safety analysts and subject matter experts, enabling rapid, accurate identification of hazardous conditions. Furthermore, this study illustrates a methodological approach for conducting NLP analysis of safety datasets that can be performed locally—isolated inside the respective organization's secure firewall, computer network, and information technology infrastructure. This element is critical to many organizations, ensuring data security and confidentiality.

How many training samples are needed for the NLP model to consistently achieve a Macro Average F_1-Score of 0.9 on a multi-class classification task?

Model performance testing metrics suggest that 60 training samples are needed for each categorical variable within the NLP classification model to achieve high precision and recall. This criterion guarantees a high degree of model performance while ensuring a manageable commitment of SME investment to categorize initial model training data.

By employing NLP methods for analyzing safety reporting narrative data, we assert that significant improvements can be realized in the classification accuracy, and timeliness of identifying and responding to potential hazards. The proposed strategy leverages the expertise of seasoned SMEs to provide initial, supervised model training that can be utilized to enhance and automate safety analysis. This efficiency enhancement, in turn, augments the capability of analysts to quickly identify potentially serious safety events. Once deployed, this tool would likely become a force multiplier within the hazard

identification process, leading to enhanced safety risk awareness, and ultimately prevention of aviation accidents.

The authors reiterate the need for new benchmarks to assess NLP model performance, particularly in areas involving high-risk datasets, including safety and accidents, healthcare, critical infrastructure, security, and legal topics. These datasets require additional care to ensure classification accuracy. Rather than relying on a single run of the NLP model to determine performance, we also advocate the importance of reporting NLP model performance using an aggregation derived from multiple runs. This approach ensures robustness against model performance variability. The authors recommend 30 run instances as an appropriate minimum in most cases. This target is generally practical and achievable while ensuring adequate statistical power for conducting analyses.

We intend to expand the number of analysis categories to determine if classification accuracy can be maintained given a limited number of training events. This is a vital step, as subject matter expert time is a scarce resource. To ensure that the tool is properly scalable for multiple events of interest, the authors need to validate that SME time spent training the model to classify events correctly can be kept to a minimum.

The authors also plan to expand this line of research to include multi-tiered analysis methods. In these methods, initial NLP filtering is conducted to correctly classify incidents into broad event categories, followed by detailed NLP analysis to identify contributing factors. This approach aims to eventually equip the model to perform root cause analysis.

Author Contributions: Conceptualization, M.D.N.; methodology, M.D.N. and R.J.W.; software, M.D.N.; validation, M.D.N. and R.J.W.; formal analysis, M.D.N. and R.J.W.; investigation, M.D.N.; resources, M.D.N. and R.J.W.; data curation, M.D.N.; writing—original draft preparation, R.J.W. and M.D.N.; writing—review and editing, M.D.N. and R.J.W.; visualization, M.D.N. and R.J.W.; supervision, M.D.N.; project administration, M.D.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This research project utilized data collected from a secondary source. The data are not publicly available, and the provider was not involved in the research project. While the provider can link the data back to living individuals, the data furnished to the authors was de-identified. Based on these factors, it was determined by the Institutional Review Board (IRB) that the research falls under 14 CFR 46 (4), Exempt Secondary Research, for which consent is not required and no IRB review is needed. The original dataset was not released because of the sensitive nature of individual safety reports and organizational safety data used in this project.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from a third party flight training provider and are available from the authors with the permission of the third party flight training provider.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tanguy, L.; Tulchecki, N.; Uriel, A.; Hermann, E.; Raynal, C. Natural language processing for aviation safety reports: From classification to interactive analysis. *Comput. Ind.* **2016**, *78*, 80–95. [CrossRef]
2. Ahadh, A.; Binish, G.V.; Srinivasan, R. Text mining of accident reports using semi-supervised keyword extraction and topic modeling. *Process Saf. Environ. Prot.* **2021**, *155*, 455–465. [CrossRef]
3. Tamascelli, N.; Paltrinieri, N.; Cozzani, V. Learning from major accidents: A meta-learning perspective. *Saf. Sci.* **2023**, *158*, 105984. [CrossRef]
4. Paraskevopoulos, G.; Pistofidis, P.; Banoutsos, G.; Georgiou, E.; Katsouros, V. Multimodal classification of safety-report observations. *Appl. Sci.* **2022**, *12*, 5781. [CrossRef]

5. Robinson, S.D. Visual representation of safety narratives. *Saf. Sci.* **2016**, *88*, 123–128. [CrossRef]
6. Oza, N.; Castle, J.P.; Stutz, J. Classification of aeronautics system health and safety documents. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2009**, *39*, 670–680. [CrossRef]
7. de Vries, V. Classification of Aviation Safety Reports using Machine Learning. In Proceedings of the 2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT), Singapore, 3–4 February 2020. [CrossRef]
8. Wolfe, S. Wordplay: An examination of semantic approaches to classify safety reports. In Proceedings of the AIAA Infotech@Aerospace 2007 Conference and Exhibit, Rohnert Park, CA, USA, 7–10 May 2007. [CrossRef]
9. Goldberg, D.M. Characterizing accident narratives with word embeddings: Improving accuracy, richness, and generalizability. *J. Saf. Res.* **2021**, *80*, 441–455. [CrossRef] [PubMed]
10. Fang, W.; Luo, H.; Xu, S.; Love, P.E.D.; Lu, Z.; Ye, C. Automated text classification of near-misses from safety reports: An improved deep learning approach. *Adv. Eng. Inform.* **2020**, *44*, 101060. [CrossRef]
11. Ricketts, J.; Barry, D.; Guo, W.; Pelham, J. A scoping literature review of natural language processing application to safety occurrence reports. *Safety* **2023**, *9*, 22. [CrossRef]
12. Horev, R. BERT explained: State of the art language model for NLP. *Towards Data Sci.* 2018. Available online: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 (accessed on 7 January 2025).
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2023**. [CrossRef]
14. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805v. [CrossRef]
15. Garrido-Merchan, E.C.; Gozalo-Brizuela, R.; Gonzalez-Carvajal, S. Comparing BERT against traditional machine learning models in text classification. *J. Comput. Cogn. Eng.* **2023**, *2*, 352–356. [CrossRef]
16. Tixier, A.J.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Autom. Constr.* **2016**, *62*, 45–56. [CrossRef]
17. Holt, R.W.; Johnson, P.J.; Goldsmith, T.E. Application of psychometrics to the calibration of air carrier evaluators. *Hum. Factors Ergon. Soc. Annu. Meet.* **1997**, *41*, 916–920. [CrossRef]
18. Yang, C.; Huang, C. Natural language processing (NLP) in aviation safety: Systematic review of research and outlook into the future. *Aerospace* **2023**, *10*, 600. [CrossRef]
19. Bex, T. Comprehensive guide to multiclass classification metrics. *Towards Data Sci.* 2021. Available online: https://towardsdatascience.com/comprehensive-guide-on-multiclass-classification-metrics-af94cfb83fbd (accessed on 7 January 2025).
20. Leung, K. Micro, macro weighted averages of F1 Score, clearly explained. *Towards Data Sci.* 2022. Available online: https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f (accessed on 7 January 2025).