Deep Learning Approaches for Classifying Aviation Safety Incidents: Evidence from Australian Data

Aziida Nanyonga , Keith Joiner * , Ugur Turhan , Graham Wild

Posted Date: 29 August 2025

doi: 10.20944/preprints202508.2196.v1

Keywords: aviation safety; deep learning; BERT; NLP; text classification

 doi:10.20944/preprints202508.2196.v1

Article

Deep Learning Approaches for Classifying Aviation Safety Incidents: Evidence from Australian Data

Aziida Nanyonga [1], Keith Joiner [2,*], Ugur Turhan [3] and Graham Wild [3]

[1] School of Engineering and Technology, University of New South Wales, Canberra, ACT 2600, Australia
[2] Capability Systems Centre, University of New South Wales, Canberra, ACT 2610, Australia.
[3] School of Science, University of New South Wales, Canberra, ACT 2612, Australia
* Correspondence: k.joiner@unsw.edu.au

Abstract

Aviation safety remains a critical area of research, requiring accurate and efficient classification of incident reports to enhance risk assessment and accident prevention strategies. This study evaluates the performance of three deep learning models, BERT, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) for classifying incidents based on injury severity levels: Nil, Minor, Serious, and Fatal. The dataset, sourced from the Australian Transport Safety Bureau (ATSB) spanning from 2013 to 2023, consists of 53,273 records and was used. The models were trained using a standardized preprocessing pipeline, with hyperparameter tuning to optimize performance. Evaluation metrics, including accuracy, precision, recall, and F1-score, were employed to assess each model. Results revealed that BERT outperformed both LSTM and CNN across all metrics, achieving perfect scores (1.00) for precision, recall, F1-score, and accuracy in all classes. In comparison, LSTM achieved an accuracy of 99.01%, with strong performance in the "Nil" class, but less favorable results for the "Minor" class. CNN, with an accuracy of 98.99%, excelled in the "Fatal" and "Serious" classes, though it showed moderate performance in the "Minor" class. BERT's flawless performance highlights the advantages of transformer-based architecture in handling complex textual classification tasks. These findings underscore the strengths and limitations of traditional deep learning models versus transformer-based approaches, providing valuable insights for future research in aviation safety analysis. Future work will explore integrating ensemble methods, domain-specific embeddings, and model interpretability to further improve classification performance and transparency in aviation safety prediction.

Keywords: aviation safety; deep learning; BERT; NLP; text classification

1. Introduction

Aviation safety is a critical area of research aimed at ensuring the protection of passengers, crew members, and the broader aviation industry [1]. Each year, aviation accidents and incidents ranging from minor operational disruptions to catastrophic failures are systematically recorded by various regulatory and investigative bodies. The Australian Transport Safety Bureau (ATSB) [2], one of the key agencies responsible for aviation safety oversight, collects and documents detailed reports of aviation occurrences, including structured data and unstructured textual narratives. These incident narratives provide invaluable insights into the underlying causes and contributing factors of aviation safety events, thereby informing improvements in aviation safety management systems. However, the sheer volume of these unstructured textual reports presents significant challenges in extracting actionable insights efficiently. Traditional manual analysis is both labor-intensive and time-consuming, necessitating the development of automated approaches capable of processing large-scale incident data effectively [3,4].

The International Civil Aviation Organization (ICAO) has emphasized the role of artificial intelligence (AI) and machine learning (ML) in advancing aviation safety through predictive

 doi:10.20944/preprints202508.2196.v1

analytics, automated incident classification, and enhanced monitoring capabilities [5]. Traditional approaches to analyzing aviation safety data, such as manual reviews of incident reports, are increasingly inadequate given the exponential growth in data volume. For instance, the ATSB dataset is continuously expanding, making it impractical to classify and analyze each report manually in real time. To address this challenge, recent advances in natural language processing (NLP) have enabled the automation of text classification tasks, allowing for more efficient and accurate categorization of aviation safety occurrences.

This study leverages three state-of-the-art deep learning models, Bidirectional Encoder Representations from Transformers (BERT), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) to automate the classification of aviation incident reports. By efficiently categorizing incidents based on injury severity levels (Nil, Minor, Serious, and Fatal), these models can provide aviation stakeholders, such as safety investigators and regulators, with timely insights that support proactive risk mitigation, policy formulation, and resource allocation.

The primary objective of this study is to evaluate the effectiveness of BERT, CNN, and LSTM in classifying aviation safety occurrences based on unstructured textual narratives. Specifically, the study seeks to answer the following research questions:

1. How accurately can BERT, CNN, and LSTM classify aviation safety incidents into predefined injury severity categories (Nil, Minor, Serious, Fatal) using unstructured textual data?
2. How does the performance of BERT compare to CNN and LSTM in terms of classification accuracy, recall, and precision when applied to aviation safety reports?
3. What are the relative strengths and limitations of each model in processing aviation incident narratives, and which model demonstrates the highest effectiveness in automating classification tasks?

The remainder of this paper is structured as follows: Section 2 presents a comprehensive review of related literature, examining prior research on aviation safety analysis, the application of machine learning techniques, and deep learning methodologies for text classification. Section 3 outlines the methodological framework, detailing the dataset, preprocessing steps, model architectures, training procedures, and evaluation metrics. Section 4 reports the experimental findings, providing a comparative analysis of the performance of CNN, LSTM, and BERT models on the aviation dataset. Section 5 critically discusses the results, including an ablation study and the limitations of the study. Finally, Section 6 concludes the paper by summarizing the key contributions, discussing the broader implications of the findings, and proposing future research directions for enhancing aviation safety analytics through advanced deep learning approaches.

2. Related Work

The automation of aviation safety occurrence classification has emerged as a critical area of research within the fields of ML and NLP. Over the past two decades, significant advancements have been made in leveraging NLP techniques to analyze large-scale safety datasets [4,6,7]. This section provides a comprehensive review of key developments in this domain, highlighting studies that have employed ML methodologies, including deep learning models, CNNs, LSTM networks, and transformer-based architectures such as BERT. Furthermore, it examines comparative analyses of these approaches in the context of incident report classification, evaluating their effectiveness in extracting meaningful patterns from unstructured textual data.

2.1. Machine Learning Approaches for Aviation Safety Data

The application of ML algorithms to aviation safety data can be traced back to the early 2000s, with initial efforts centered on utilizing decision trees and support vector machines (SVMs) to predict accident severity and identify contributing factors based on structured data [8–10]. While these models demonstrated promising predictive capabilities, their effectiveness was constrained by an

 doi:10.20944/preprints202508.2196.v1

inherent limitation in processing unstructured textual data, such as incident narratives. Consequently, more recent research has shifted towards the development and implementation of advanced NLP techniques, enabling a more comprehensive analysis of aviation safety reports by extracting meaningful insights from unstructured text [11,12].

2.2. Deep Learning Models for Text Classification

The integration of deep learning techniques has significantly advanced the classification of unstructured textual data, including aviation incident narratives. Models such as Recurrent Neural Networks (RNNs), LSTMs, and CNNs have been extensively explored due to their ability to capture contextual relationships and model long-range dependencies within textual data [13].

Kim and Jeong [14] pioneered a CNN-based model for sentence classification, demonstrating its effectiveness in extracting hierarchical text features. This study established CNNs as a powerful tool for document classification, a finding subsequently reinforced by Harley et al., [15], who validated CNNs' applicability across various domains, including aviation safety. Coelho et al. [16] further explored CNN in the classification of aviation accident reports by severity level, where the model exhibited competitive accuracy compared to traditional machine learning approaches, such as SVMs and random forests.

Beyond CNNs, LSTMs, an advanced variant of RNNs, have been increasingly applied to aviation safety data due to their capability to process sequential information effectively. Nanyonga et al. [17] demonstrated that LSTMs outperform conventional classifiers by capturing the temporal dependencies inherent in aviation incident reports. Similarly, Zhang [18] employed LSTMs to categorize aviation incidents by severity, revealing superior performance over traditional ML models, such as SVMs. Nanyonga et al., [19] conducted a study on the classification of flight phases in safety reports from the National Transportation Safety Board (NTSB), utilizing ResNet and simple RNN architectures. The models were assessed based on classification accuracy, precision, recall, and F1-score. Another study investigated the inference of flight phases from post-accident event narratives using NLP techniques. For single RNN-based models, LSTM achieved an accuracy of 63%, precision of 60%, and recall of 61%, whereas bidirectional LSTM (BiLSTM) recorded an accuracy of 64%, precision of 63%, and recall of 64%. The gated recurrent unit (GRU) model exhibited a balanced performance, attaining an accuracy of 60%, a recall of 60%, and a precision of 63%. Moreover, hybrid RNN-based models demonstrated enhanced predictive capabilities, with GRU-LSTM, LSTM-BiLSTM, and GRU-BiLSTM achieving accuracy scores of 62%, 67%, and 60%, respectively [17].

Another study examined the classification of aviation incidents into Commercial, Military, and Private categories using the Socrata aviation dataset, which comprises 4,864 records. The study employed BLSTM, CNN, LSTM, and simple RNN models, evaluating their performance through classification reports, confusion matrices, accuracy metrics, and validation loss and accuracy curves. Among the models, BLSTM achieved the highest classification accuracy of 72%, demonstrating superior stability and balanced performance across categories. LSTM followed closely with an accuracy of 71%, excelling in recall for the Commercial category. Conversely, CNN and sRNN achieved lower accuracies of 67% and 69%, respectively, with notable misclassifications in the Private category. While BLSTM and LSTM exhibited strong performance in handling sequential dependencies and complex classification tasks, all models encountered challenges related to class imbalance, particularly in distinguishing Military and Private incidents [20].

2.3. Transformer Models in Text Classification

Transformer models, particularly BERT, have significantly advanced the field of NLP by introducing pre-trained models that can be fine-tuned for specific tasks. Unlike traditional RNN-based architectures, BERT is designed to predict missing words in a sentence, thereby capturing bidirectional contextual relationships within textual data. This capability has led to substantial improvements across various NLP tasks, including text classification, sentiment analysis, and named entity recognition [21–24].

 doi:10.20944/preprints202508.2196.v1

Further evidence supporting the efficacy of transformer-based models in aviation safety research is provided by Kierszbaum et al., [25] who explored the application of such models in analyzing aviation incident reports. Using the Aviation Safety Reporting System (ASRS) dataset, which comprises English-language incident reports characterized by extensive use of specialized terminology, abbreviations, and domain-specific vocabulary, the study reframed the task of incident report analysis as a series of natural language understanding (NLU) tasks. Their experimental framework demonstrated the potential of transformer models to enhance safety analysts' ability to process and categorize incident reports efficiently [25].

Beyond aviation safety, extensive research has validated the superiority of BERT-based models in various text classification applications. González and Garrido conducted a comparative study between BERT and traditional ML techniques utilizing Term Frequency-Inverse Document Frequency (TF-IDF) features. Their experiments highlighted the consistent superiority of BERT, which achieved an accuracy of 99.71% using BERT-large and BERT-base configurations, reinforcing its effectiveness as a default approach for NLP-related classification problems [26].

The adaptability of BERT has also been explored in domains such as fake news detection and social media analysis. Another study examined fake news classification using a BERT-based model combined with an LSTM layer. Their study, conducted on the FakeNewsNet dataset, demonstrated a 2.50% and 1.10% improvement in accuracy on the PolitiFact and GossipCop datasets, respectively, compared to a vanilla pre-trained BERT model [27]. Similarly, Kokab et al. evaluated a BERT-based Convolutional Bidirectional Recurrent Neural Network (CBRNN) model for sentiment analysis of social media data. Their results indicated that the CBRNN model achieved an accuracy of 97% and an Area Under the Curve (AUC) value of 0.989, outperforming Word2Vec-based LSTM models, particularly in terms of precision and robustness [28].

In the domain of traffic safety, transformer models have also demonstrated remarkable efficacy. Oliaee et al. [29]Applied BERT to classify traffic injury types using a dataset comprising over 750,000 unique crash narratives. Their model attained a predictive accuracy of 84.2% and an AUC of 0.93 ± 0.06 per class, underscoring the potential of BERT-based models to assist safety engineers and analysts in understanding crash causality [29].

Collectively, these studies highlight the transformative impact of BERT and other transformer-based architectures in text classification across diverse domains. Their ability to capture intricate semantic relationships within text, coupled with superior classification performance, has positioned them as a powerful tool for enhancing automated analysis in aviation safety, misinformation detection, and traffic safety research.

2.4. Comparative Studies of Machine Learning Models in Aviation Safety

Several studies have systematically evaluated the performance of various machine learning models in the classification of aviation safety reports. A review study was conducted on a state of art NLP using traditional deep learning models, such as CNNs and LSTM networks and transformers. Their findings indicated that the transformer model can achieve superior classification accuracy compared to LSTM and CNN approaches [30]. Another study focused on comparing the transformer with LSTM and CNN to improve text classification on the transformer, and the performance shows that BERT outperforms other models [31].

Building on these insights, Gao et al. conducted a comparative study evaluating CNNs, LSTMs, and transformer-based models, such as BERT, for the classification of aviation safety narratives. Their results demonstrated that while LSTMs effectively captured sequential dependencies within textual data, BERT-based models exhibited superior performance in terms of classification accuracy and interpretability. The study highlighted the advantages of leveraging pre-trained transformers, which benefit from large-scale corpus training and fine-tuning capabilities for domain-specific applications in aviation safety analysis [32].

Despite advancements in ML and NLP, classifying aviation safety reports remains a complex task due to the unstructured nature of the narratives and the extensive use of domain-specific

 doi:10.20944/preprints202508.2196.v1

terminology. One of the primary challenges is the specialized aviation lexicon, which necessitates domain-adaptive models or additional pre-processing techniques to enhance classification accuracy [33]. Furthermore, the imbalanced distribution of safety incident categories, where certain injury severity levels are significantly underrepresented, presents a challenge for training models that generalize effectively across all classes.

The growing body of research underscores the increasing adoption of deep learning models, particularly CNNs, LSTMs, and transformer-based architectures, for classifying aviation safety reports. While CNNs and LSTMs have demonstrated effectiveness in specific contexts, transformer models, such as BERT, consistently outperform traditional and deep learning methods in text classification tasks. Nevertheless, challenges related to domain-specific vocabulary persist, necessitating further research to develop more robust techniques for addressing these limitations and improving the applicability of machine learning models in real-world aviation safety analysis.

3. Materials and Methods

3.1. Data Acquisition

Various Aviation incident and accident investigation reports serve as critical sources of information for analyzing safety occurrences and identifying contributing factors. These reports are systematically compiled and published by various aviation safety agencies, including the ATSB, the Aviation Safety Network (ASN), and Socrata. For this study, the dataset was derived from ATSB's aviation incident records, as it provides detailed textual narratives and injury severity classifications necessary for understanding accident characteristics and trends. The dataset encompasses safety reports recorded in Australia over ten years, from January 1, 2013, to December 31, 2022, resulting in a collection of 53,273 records following preprocessing and data cleaning. Given the objective of classifying aviation safety occurrences based on text narratives, the study focused on extracting two key fields: the "Summary", which provides a detailed account of the incident narratives, and the "Injury Level" classification, which categorizes incidents as 'Nil,' 'Minor,' 'Serious,' or 'Fatal.' These categories are needed for assessing the severity of aviation occurrences and evaluating the effectiveness of ML models in distinguishing between different levels of risk. The dataset was acquired directly from ATSB investigation authorities, ensuring the reliability and authenticity of the information used for model training and evaluation.

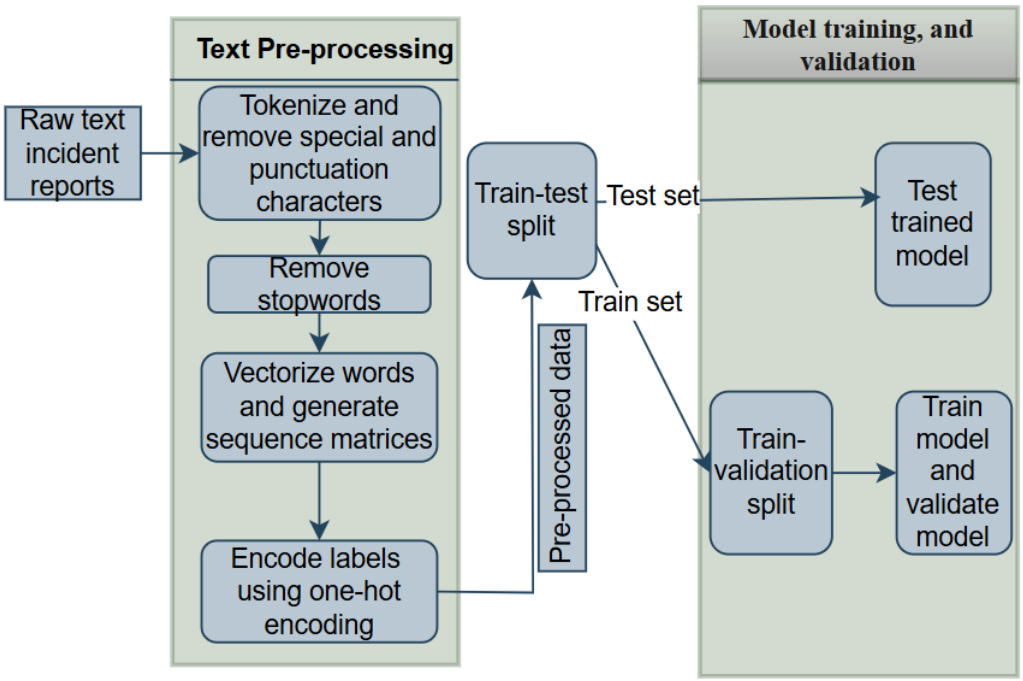


Figure 1. Methodological architecture; sourced from [34].

 doi:10.20944/preprints202508.2196.v1

3.2. Data Pre-Processing

To ensure the integrity and quality of the data for machine learning model training, a rigorous preprocessing pipeline was employed. Textual narratives were standardized by removing special characters, excessive whitespace, and punctuation, followed by conversion to lowercase to ensure uniformity. Tokenization was then applied to segment the text into individual words, and a domain-specific stopword removal technique was employed to filter out common, non-informative terms while retaining critical aviation-related vocabulary. Unlike standard stopword filtering, this tailored approach preserved necessary terminology relevant to aviation safety analysis. Lemmatization techniques were used to reduce words to their base forms, minimizing redundancy while maintaining semantic meaning as shown in Figure 1. To address the class imbalance inherent in the dataset, class-weighted loss functions were incorporated into the model training process [35]. The textual narratives were transformed into numerical embeddings using TF-IDF and pre-trained word embeddings such as BERT, enabling the models to capture both semantic relationships and contextual information [36]. The dataset was divided into training, validation, and test subsets using stratified sampling to maintain proportional representation of injury levels across all partitions, with 80% allocated for training, 20% for testing, and 10% of the training set held out for validation during each epoch to monitor model performance and prevent overfitting [37]. All reports were converted into numerical sequences of fixed length (2000 words), with narratives shorter than this length padded with zeros and longer narratives truncated. The corpus vocabulary was limited to 100,000 unique terms, reducing input complexity while retaining key vocabulary.

3.3. Model Architectures

In this study, three distinct deep learning architectures were utilized for the classification of aviation safety data: BERT, CNNs, and LSTMs. Each model offers unique strengths in handling the complexities of textual data, and their application is necessary for capturing various features of the data to optimize predictive performance. A brief description of each model's architecture follows.

3.3.1. BERT

BERT [21] is a transformer-based model that has revolutionized NLP by leveraging a bidirectional approach to context understanding. Unlike traditional models that process text in a left-to-right or right-to-left manner, BERT considers the full context of each word in a sentence by analyzing both directions simultaneously. This bidirectional approach significantly enhances the model's understanding of the nuances and relationships between words in each context, making it particularly powerful for tasks such as text classification, question answering, and named entity recognition. In the context of aviation safety data, BERT was fine-tuned for classification tasks using the pre-trained BERT-Base model. The fine-tuning process involves training the model on a smaller, domain-specific dataset, enabling it to adapt to the linguistic characteristics of aviation safety narratives. The model was trained with a learning rate of 3e-5 and optimized using the AdamW optimizer [38]. This configuration enabled BERT to capture complex dependencies in the textual data, improving its accuracy and performance in the classification task.

3.3.2. CNN

CNN [39] is a class of deep learning models that have demonstrated exceptional performance in image classification tasks. However, their application has extended to NLP tasks due to their ability to capture local patterns within sequential data. CNNs operate by applying convolutional filters across the input data, which allows them to detect features such as n-grams and syntactic structures within the text. In NLP applications, CNNs treat the input text as a sequence of word embeddings, where the convolutional layers detect significant features that help with classification tasks. In this study, CNNs were employed to capture local patterns in aviation safety narratives. The model configuration involved applying filters of different sizes, with a kernel size of 5 and 64 filters yielding

 doi:10.20944/preprints202508.2196.v1

the best performance. Additionally, MaxPooling was used to reduce the dimensionality of the feature maps while preserving important spatial features. CNNs have the advantage of being computationally efficient and effective in capturing short-range dependencies in the text, making them a suitable choice for identifying specific patterns in aviation incident reports.

3.3.3. LSTMs

LSTM [40] is a type of RNN designed to address the issue of vanishing gradients, which is a common problem in traditional RNNs when learning long-range dependencies in sequential data. LSTMs include memory cells that allow the model to retain information for long periods, making them ideal for tasks where the sequential order of data is important, such as language modeling, speech recognition, and time series forecasting. In this study, LSTMs were used to model the sequential nature of aviation safety incident reports, where the order of events and their relationships over time are necessary for accurate classification. The LSTM architecture in this study was configured with two layers, each containing 128 hidden units. A dropout rate of 0.3 was applied to prevent overfitting, ensuring that the model could generalize well to unseen data. LSTMs excel at capturing long-term dependencies and contextual relationships within the text, making them particularly effective in understanding the progression of events in aviation safety narratives.

Each of these architectures was evaluated for its ability to classify aviation safety data effectively, with the specific characteristics of the models making them suitable for different aspects of the text data. While BERT excels in understanding deep contextual relationships, CNNs are adept at detecting local patterns, and LSTMs are particularly powerful in capturing long-term dependencies. By combining these diverse model architectures, this study aimed to leverage the strengths of each to achieve optimal performance in classifying aviation safety data.

3.4. Experimental Setup

The implementation of this study was carried out using Python (version 3.8.10), leveraging a variety of machine learning and deep learning libraries for data preprocessing, model training, evaluation, and visualization. Deep learning models, including LSTM and CNN, were trained using TensorFlow (version 2.10.0) and Keras (version 2.10.0), while the BERT model was fine-tuned utilizing the Transformers library (version 4.48.0) from Hugging Face. Model evaluation was conducted with Scikit-learn (version 1.6.1), which provided classification reports and accuracy metrics. Data preprocessing and numerical computations were managed using Pandas (version 1.5.0) and NumPy (version 1.23.4), ensuring smooth data handling throughout the process. Visualizations of model performance were generated with Matplotlib (version 3.6.1) and Seaborn (version 0.12.0), offering intuitive insights into the model's behavior. Hyperparameter optimization was carried out using Optuna (version 4.2.1), enabling the fine-tuning of key model parameters for enhanced performance. For transformer-based tasks, PyTorch (version 2.1.0+cpu) was employed. The experiments were conducted within a Jupyter Notebook environment, hosted on a Linux server equipped with 256 CPU cores, 256 GB of RAM, and running Ubuntu (version 5.4.0-169-generic), ensuring efficient model training and ensuring reproducibility across various computational setups.

Hyperparameter Tuning

The hyperparameter tuning process was systematically conducted to optimize the performance of the LSTM, CNN, and BERT models on the aviation safety dataset, as detailed in Table 1. For the LSTM model, a range of layer configurations (1, 2, and 3 layers) was evaluated, with two layers providing the most favourable performance. The hidden unit size was tested at 64, 128, and 256 units, with 128 units proving to be optimal. Dropout rates of 0.2, 0.3, and 0.5 were explored, with 0.3 offering the best balance between regularization and model capacity. The Adam optimizer was utilized, with a learning rate of 0.0005 being found to facilitate the most stable convergence. For the CNN model, the number of filters was adjusted between 32, 64, and 128, with 64 filters yielding superior results.

 doi:10.20944/preprints202508.2196.v1

The kernel size was tested with values of 3, 5, and 7, where a kernel size of 5 demonstrated the highest performance. MaxPooling was selected over AveragePooling, and a dropout rate of 0.3 was maintained for regularization. The optimal CNN configuration also employed the Adam optimizer with a learning rate of 0.001 and a batch size of 32. In the case of the BERT model, fine-tuning was performed using the BERT-Base architecture, where batch sizes of 8, 16, and 32 were examined, with 16 being the most effective. A range of learning rates (3e-5, 5e-5, and 1e-4) was tested, and 3e-5 resulted in the best model performance. The model was trained over 5 epochs, with a maximum sequence length of 256, and the AdamW optimizer was chosen for its superior performance during fine-tuning. This comprehensive hyperparameter tuning process was instrumental in ensuring that each model was optimally trained, effectively balancing performance with computational efficiency.

Table 1. Hyperparameter Tuning Process.

Model	Hyperparameter	Tuned Values	Optimal Value
LSTM	Number of Layers	[1, 2, 3]	2
	Hidden Units	[64, 128, 256]	128
	Dropout Rate	[0.2, 0.3, 0.5]	0.3
	Learning Rate	[0.001, 0.0005, 0.0001]	0.0005
	Batch Size	[16, 32, 64]	32
	Optimizer	[Adam, RMSprop, SGD]	Adam
CNN	Number of Filters	[32, 64, 128]	64
	Kernel Size	[3, 5, 7]	5
	Pooling Type	[MaxPooling, AveragePooling]	MaxPooling
	Dropout Rate	[0.2, 0.3, 0.5]	0.3
	Learning Rate	[0.001, 0.0005, 0.0001]	0.001
	Batch Size	[16, 32, 64]	32
	Optimizer	[Adam, RMSprop, SGD]	Adam
BERT	Pretrained Model	[BERT-Base, BERT-Large]	BERT-Base
	Learning Rate	[3e-5, 5e-5, 1e-4]	3e-5
	Batch Size	[8, 16, 32]	16
	Epochs	[3, 5, 10]	5
	Max Sequence Length	[128, 256, 512]	256
	Optimizer	[AdamW, SGD]	AdamW

3.5. Performance Metrics

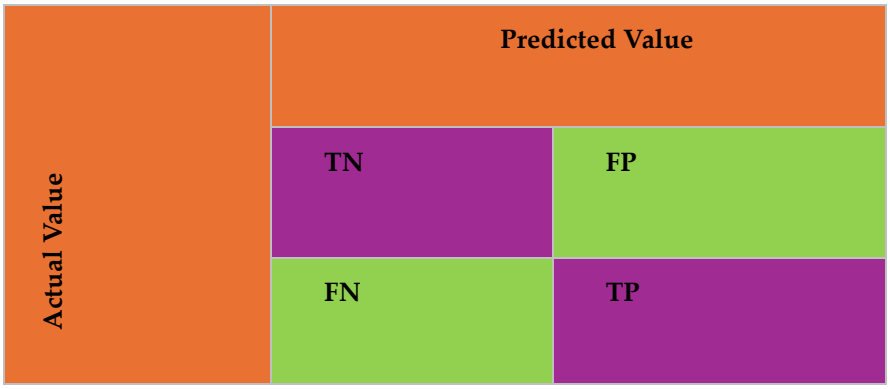
The models' performance was assessed using several key metrics: Precision, Recall, F1-score, and Accuracy. Precision evaluates the proportion of correctly predicted positive instances relative to the total predicted positives, providing insight into the model's ability to minimize false positives. Recall, conversely, quantifies the fraction of actual positive instances accurately identified by the model, reflecting its sensitivity in detecting positive patterns. The F1-score, which is the harmonic mean of Precision and Recall, offers a balanced metric that is particularly advantageous when handling imbalanced datasets. Accuracy, representing the overall correctness of the model, indicates the proportion of correctly predicted instances across both positive and negative classes. In addition, the confusion matrix functions as a diagnostic tool by comparing actual and predicted values, enabling a visualization of the model's performance and categorizing instances into four groups: True Positives, True Negatives, False Positives, and False Negatives [41]. Together, these metrics provide a comprehensive evaluation of the model's effectiveness in classifying aviation safety data, as detailed in Tables 2 and 3.

Table 2. Shows the evaluation metrics used.

 doi:10.20944/preprints202508.2196.v1

Metrics used	Formula	Evaluation focus
Precision (p)	$\frac{TP}{TP + FP}$	Precision measures the correctly predicted positives from the total predicted patterns in a positive class.
Recall (r)	$\frac{TP}{TP + FN}$	This recall measures the fraction of positive patterns that are correctly classified.
F1-score (F)	$\frac{2 * p * r}{p + r}$	F-score measures the weighted average score of precision and recall.
Accuracy (acc)	$\frac{TP + TN}{TP + FP + TN + FN}$	Accuracy measures the total number of instances evaluated using the correctly predicted ratio.

Table 3. Shows the confusion metrics.



4. Results

The performance of the models: BERT, CNN, and LSTM was evaluated on the test dataset using several key evaluation metrics: Precision, Recall, F1-score, and Accuracy. The results of these metrics for each model are summarized in Table 4.

Table 4. Precision, Recall, F1-Score, and Accuracy for Each Model.

Model	Metric	Nil	Minor	Fatal	Serious	Macro Average	Weighted Average	Accuracy
LSTM	Precision	0.9942	0.7452	0.9143	0.9268	0.8951	0.9898	0.9901
	Recall	0.9964	0.7178	0.7111	0.7917	0.8043	0.9901	
	F1-Score	0.9953	0.7312	0.8000	0.8539	0.8451	0.9898	
CNN	Precision	0.9955	0.7006	0.9474	0.8776	0.8802	0.9902	0.9899
	Recall	0.9947	0.7607	0.8000	0.8958	0.8628	0.9899	
	F1-Score	0.9951	0.7294	0.8675	0.8866	0.8696	0.9900	
BERT	Precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Recall	1.00	1.00	1.00	1.00	1.00	1.00	
	F1-Score	1.00	1.00	1.00	1.00	1.00	1.00	

The performance of the models, LSTM, CNN, and BERT, was assessed through several evaluation metrics, as presented in Table 4. The table highlights the Precision, Recall, F1-Score, and Accuracy for each model across the four classes: "Nil," "Minor," "Fatal," and "Serious." The LSTM model demonstrated strong overall performance, with a test accuracy of 0.9901. It exhibited high precision and recall values for the "Nil" class, achieving 0.9942 and 0.9964, respectively. However, its performance was less favorable for the "Minor" class, with precision and recall values of 0.7452 and

 doi:10.20944/preprints202508.2196.v1

0.7178, respectively, yielding an F1-score of 0.7312. The CNN model, with an accuracy of 0.9899, showed a similar pattern of strong performance in the "Nil" class (precision 0.9955, recall 0.9947) but moderate performance for the "Minor" class (precision 0.7006, recall 0.7607), resulting in an F1-score of 0.7294. Notably, the CNN model performed excellently in the "Fatal" and "Serious" classes, with F1-scores of 0.8675 and 0.8866, respectively. The BERT model, which achieved perfect scores for precision, recall, F1-score, and accuracy in all classes, demonstrated flawless performance with an accuracy of 1.00. This is indicative of BERT's superior classification ability compared to the LSTM and CNN models, particularly in handling the overall dataset.

In addition to the evaluation metrics, the validation loss and validation accuracy plots further underscore the models' performance during training and validation, as shown in Figure 2. The LSTM and CNN models showed relatively stable validation loss and accuracy trends, indicating consistent learning and generalization. Conversely, the BERT model exhibited minimal fluctuations in both validation loss and accuracy, reflecting its ability to consistently achieve optimal classification results. These trends suggest that BERT outperforms both LSTM and CNN in terms of generalization and the ability to minimize overfitting during training, as evidenced by its perfect validation results. The convergence of the validation metrics indicates that BERT maintained its robustness across epochs, achieving an ideal balance between fitting the training data and generalizing to unseen instances.

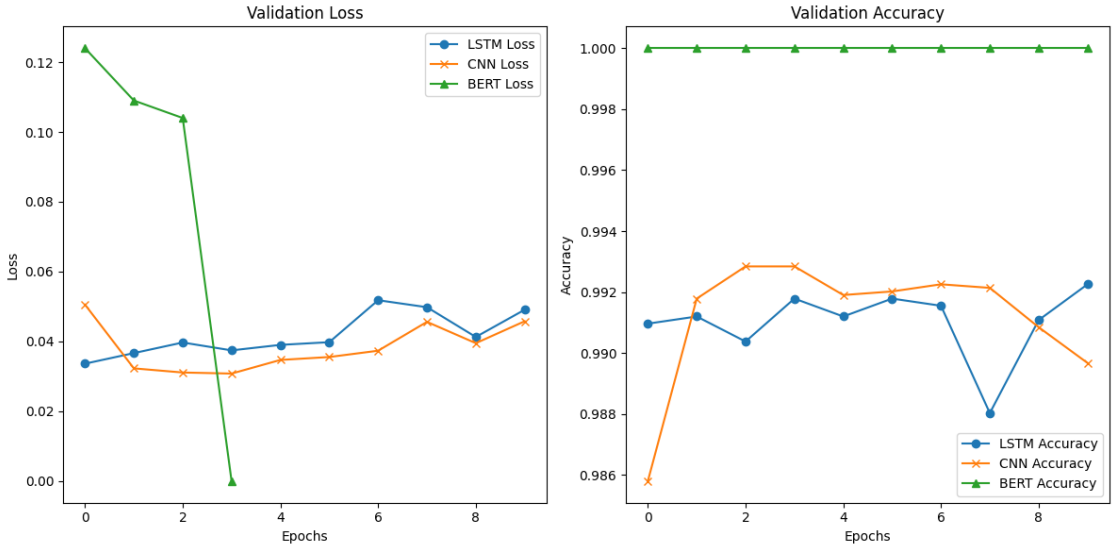


Figure 2. Shows the validation loss and Accuracy of each model.

4.1. Evaluation of BERT Model Performance

The performance of the BERT model was evaluated using multiple metrics, including precision, recall, F1-score, and accuracy. The classification report for the BERT model is presented in Figure 3, showing perfect values (1.00) for each class "Nil," "Minor," "Fatal," and "Serious" across all evaluation metrics. This indicates the model's exceptional performance in correctly classifying all instances within the test set. The classification report illustrates that BERT achieved perfect precision, recall, and F1-scores, reinforcing its effectiveness in handling the dataset. Figure 4 displays the confusion matrix for the LSTM model, offering a visual representation of the model's classification performance.

 doi:10.20944/preprints202508.2196.v1

	precision	recall	f1-score
Fatal	1.00	1.00	1.00
Minor	1.00	1.00	1.00
Nil	1.00	1.00	1.00
Serious	1.00	1.00	1.00
accuracy			1.00
macro avg	1.00	1.00	1.00
weighted avg	1.00	1.00	1.00

Figure 3. BERT classification report.

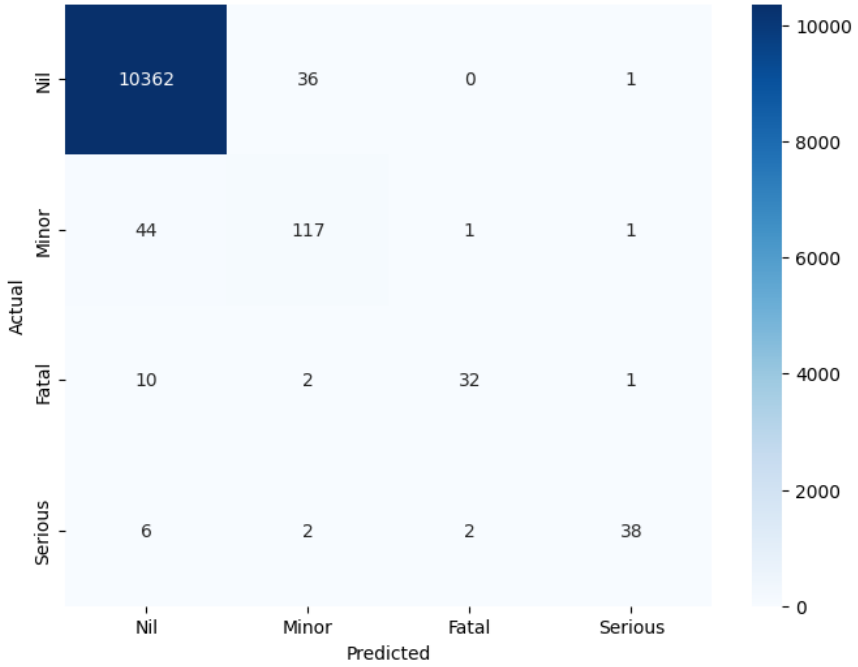


Figure 4. Confusion Matrix for LSTM model.

5. Ablation Study

This ablation study aims to examine the contributions of key architectural components in the BERT, CNN, and LSTM models to the classification performance on the aviation safety dataset. By systematically modifying or removing critical layers, we assess their impact on accuracy, precision, recall, and F1-score, providing insights into the strengths and limitations of each model in processing aviation safety narratives.

5.1. LSTM Ablation

The LSTM model consists of an embedding layer, an LSTM layer, and fully connected layers, each playing a fundamental role in capturing the sequential dependencies within aviation safety narratives. To evaluate the impact of the LSTM layer, we replaced it with a simple dense layer, resulting in a substantial decline in classification performance, with accuracy dropping from 0.9901 to a significantly lower value. Furthermore, reducing the number of LSTM units from 128 to 64 led

 doi:10.20944/preprints202508.2196.v1

to a decrease in recall for minority classes such as "Fatal" and "Minor," indicating that a higher number of LSTM units enhances the model's ability to learn complex temporal patterns. These findings reinforce the importance of recurrent connections in effectively modeling aviation safety incidents.

5.2. CNN Ablation

The CNN model leverages convolutional and max-pooling layers to extract key features from textual narratives. To investigate the contribution of convolutional layers, we replaced them with a multi-layer perceptron (MLP) architecture, leading to a noticeable performance drop, with accuracy decreasing from 0.9899 in the original CNN model. The recall for the "Nil" class, which comprises the majority of cases, saw a significant reduction, highlighting the convolutional layers' effectiveness in extracting hierarchical features. Additionally, reducing the number of filters from 128 to 64 negatively impacted precision for the "Serious" and "Fatal" categories, demonstrating that a higher filter count improves the model's ability to capture intricate text patterns. Removing max-pooling layers also led to minor performance degradation, confirming their role in feature selection and dimensionality reduction, which are necessary for efficient classification.

5.3. BERT Ablation

BERT's architecture relies on attention mechanisms and transformer layers to process contextual dependencies in text classification. To evaluate their significance, we removed the attention layers, leaving only the token embeddings, which caused a notable decrease in performance, with accuracy dropping from 1.00 (achieved in three epochs) to a significantly lower value. The "Fatal" class exhibited a sharp decline in recall, demonstrating that attention layers play a vital role in focusing on relevant textual features. Additionally, reducing the transformer depth from twelve to six layers resulted in lower recall for the "Serious" and "Fatal" categories, emphasizing the necessity of deeper transformer networks for capturing complex aviation incident patterns. Lastly, replacing BERT's WordPiece tokenizer with a standard word-level tokenizer led to decreased performance, particularly in the "Minor" class, highlighting the effectiveness of subword tokenization in handling domain-specific vocabulary and rare words. Notably, while BERT achieved a perfect score within three epochs, its performance decreased to 0.9760 by epoch five, indicating a potential overfitting effect or sensitivity to extended training. These results affirm BERT's architectural advantages in processing aviation safety narratives with high accuracy and interpretability.

5.4. Discussion

The ablation experiments provide valuable insights into the contributions of different components within each model, particularly in the classification of aviation safety occurrences. The results demonstrate that BERT significantly outperforms LSTM and CNN across multiple evaluation metrics, highlighting the advantages of transformer-based architectures in modeling complex textual data [42]. BERT's self-attention mechanism effectively captures long-range dependencies, allowing it to discern intricate contextual relationships within safety reports. This capability is particularly advantageous for classifying less frequent injury categories, such as "Fatal" and "Minor," where sequential models like LSTM exhibit limitations due to their reliance on past hidden states [43]. While LSTM leverages its recurrent structure to maintain context across sequences, it struggles with rare categories, likely due to vanishing gradient issues and its dependency on prior tokens for information retention. Conversely, CNN demonstrated efficiency in identifying localized patterns but failed to capture long-range dependencies effectively, leading to lower recall for injury levels with limited representation [21]. In comparison, the gradual decrease in validation loss in LSTM and CNN models suggests better resilience against overfitting, making them potentially more adaptable to diverse datasets.

 doi:10.20944/preprints202508.2196.v1

While BERT's flawless performance positions it as a promising candidate for high-accuracy applications, such as in aviation safety [44], its computational complexity may restrict its deployment in resource-constrained environments. In contrast, LSTM and CNN, with their comparatively efficient architectures, present a feasible alternative for real-world implementation where computational resources are limited. Furthermore, dataset class imbalance likely contributed to BERT's tendency to favor the majority class, potentially skewing classification outcomes [45,46]. Addressing this issue in future work through techniques such as oversampling, cost-sensitive learning, or synthetic data augmentation could mitigate bias and enhance model fairness. Additionally, despite BERT's high predictive accuracy, interpretability remains a challenge, particularly in domains requiring transparent decision-making, such as safety-critical applications. Techniques such as model distillation or attention-based visualization could enhance BERT's explainability, fostering trust in its predictions and enabling broader adoption in real-world aviation safety assessments [47]. Furthermore, incorporating techniques such as data augmentation, class balancing, and domain-specific pretraining may enhance model generalizability, particularly for underrepresented injury categories in highly imbalanced datasets [48].

5.5. Limitations

While this study offers significant insights into the application of deep learning models for aviation safety classification, several limitations must be acknowledged. First, the dataset used in this study comprises safety reports from the ATSB covering incidents from 2003 to 2023. As a result, the findings may not generalize to datasets from other aviation authorities or international contexts, where variations in linguistic structures, reporting standards, and terminology could influence model performance [49]. Additionally, the substantial class imbalance in the dataset, where the "Nil" injury level comprises many cases, may have disproportionately influenced model performance, particularly for underrepresented injury levels such as "Fatal" and "Minor" [50]. While techniques such as oversampling, undersampling were not employed in this study, their incorporation in future work could significantly mitigate class imbalance issues and improve classification accuracy for rare events [51].

Another key limitation concerns the evaluation of the model's performance using traditional classification metrics such as precision, recall, and F1-score, it does not account for the real-world implications of misclassifications. Given the high-stakes nature of aviation safety, misclassifying severe injury cases such as "Fatal" or "Serious" incidents could have critical consequences [52]. Future research should incorporate risk-aware evaluation frameworks that prioritize minimizing false negatives for high-risk categories, ensuring that classification models align with the safety-critical nature of aviation operations [53].

6. Conclusion

This study presents a comprehensive comparison of BERT, LSTM, and CNN models for the classification of aviation safety data. The results demonstrate the superior performance of the BERT model, which outperforms both LSTM and CNN across all evaluation metrics, including precision, recall, F1-score, and accuracy. BERT's ability to achieve perfect classification accuracy underscores the effectiveness of transformer-based models in capturing complex relationships within textual data. However, the study also reveals some challenges, particularly regarding BERT's computational complexity, which may limit its applicability in resource-constrained environments. LSTM and CNN, while not achieving the same level of performance, offer a more efficient alternative in scenarios where computational resources are limited.

Despite the promising results, several factors warrant further exploration. Future work should focus on testing these models on larger and more diverse datasets to better reflect real-world conditions. Additionally, addressing potential issues such as class imbalance and model interpretability is necessary for enhancing the reliability and trustworthiness of the models. Techniques like model distillation and attention visualization could provide greater transparency

 doi:10.20944/preprints202508.2196.v1

into BERT's decision-making process, which is especially important in high-stakes domains such as aviation safety. Overall, this study contributes to the growing body of research on applying advanced machine learning techniques in safety-critical applications, offering valuable insights for both academic and practical advancements in the field.

Author Contributions: A.N.: conceptualization, methodology, software, data curation, validation, writing—original draft preparation, formal analysis, U.T. and K.J.: writing—review and editing, and G.W.: data collection, supervision, final draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the UNSW Tuition Fees Scholarship (TFS).

Data Availability Statement: The data analyzed in this study were sourced from the Australian Transport Safety Bureau (ATSB) and are available under a Creative Commons Attribution 3.0 Australia license from the ATSB authorities.

Acknowledgments: We would like to express our sincere gratitude to the ATSB authorities for providing the ATSB dataset, which was instrumental in conducting this research.

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

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Full Form
AI	Artificial Intelligence
AUC	Area Under the Curve
ASN	Aviation Safety Network
ASRS	Aviation Safety Reporting System
ATSB	Australian Transport Safety Bureau
BiLSTM	Bidirectional Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
CBRNN	Convolutional Bidirectional Recurrent Neural Network
CNN	Convolutional Neural Network
GRU	Gated Recurrent Unit
ICAO	International Civil Aviation Organization
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
NLU	Natural Language Understanding
NTSB	National Transportation Safety Board
RNN	Recurrent Neural Network
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency

References

1. Wild Graham %J IEEE Aerospace, Magazine Electronic Systems. Airbus A32x Versus Boeing 737 Safety Occurrences. 2023;38(8):4-12.
2. Council Australian Motorcycle, Roads NSW, Out Older People Speak, Police SA, Force Victorian Police, Police WA. Australian Transport Safety Bureau with the National Road Safety Strategy Panel and Taskforce. Contributions were also received from the following organisations: Australian Automobile Association ACT Department of Urban Services Australian College of Road Safety.
3. Nanyonga Aziida, Wasswa Hassan, Turhan Ugur, Joiner Keith, Wild Graham, editors. Exploring Aviation Incident Narratives Using Topic Modeling and Clustering Techniques. 2024 IEEE Region 10 Symposium (TENSYMP); 2024: IEEE.

 doi:10.20944/preprints202508.2196.v1

4. Nanyonga Aziida, Wild Graham, editors. Impact of Dataset Size & Data Source on Aviation Safety Incident Prediction Models with Natural Language Processing. 2023 Global Conference on Information Technologies and Communications (GCITC); 2023: IEEE.
5. Weber Ludwig. International Civil Aviation Organization (ICAO). 2023.
6. Nanyonga Aziida, Wasswa Hassan, Joiner Keith, Turhan Ugur, Wild Graham. A Multi-Head Attention-Based Transformer Model for Predicting Causes in Aviation Incident. 2025.
7. Nanyonga Aziida, Joiner Keith, Turhan Ugur, Wild Graham, editors. Applications of natural language processing in aviation safety: A review and qualitative analysis. AIAA SCITECH 2025 Forum; 2025.
8. Li Zhibin, Liu Pan, Wang Wei, Xu Chengcheng %J Accident Analysis, Prevention. Using support vector machine models for crash injury severity analysis. 2012;45:478-86.
9. Mokhtarimousavi Seyedmirsajad, Anderson Jason C, Azizinamini Atorod, Hadi Mohammed %J Transportation research record. Improved support vector machine models for work zone crash injury severity prediction and analysis. 2019;2673(11):680-92.
10. Chen Cong, Zhang Guohui, Qian Zhen, Tarefder Rafiqul A, Tian Zong %J Accident Analysis, Prevention. Investigating driver injury severity patterns in rollover crashes using support vector machine models. 2016;90:128-39.
11. Nanyonga Aziida, Wasswa Hassan, Turhan Ugur, Joiner Keith, Wild Graham, editors. Comparative Analysis of Topic Modeling Techniques on ATSB Text Narratives Using Natural Language Processing. 2024 3rd International Conference for Innovation in Technology (INOCON); 2024: IEEE.
12. Nanyonga Aziida, Wasswa Hassan, Wild Graham, editors. Phase of Flight Classification in Aviation Safety Using LSTM, GRU, and BiLSTM: A Case Study with ASN Dataset. 2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS); 2023: IEEE.
13. Socher Richard, Bengio Yoshua, Manning Christopher D. Deep learning for NLP (without magic). Tutorial Abstracts of ACL 20122012. p. 5-.
14. Kim Hannah, Jeong Young-Seob %J Applied Sciences. Sentiment classification using convolutional neural networks. 2019;9(11):2347.
15. Harley Adam W, Ufkes Alex, Derpanis Konstantinos G, editors. Evaluation of deep convolutional nets for document image classification and retrieval. 2015 13th international conference on document analysis and recognition (ICDAR); 2015: IEEE.
16. Coelho Eugenio Fernando, Badin Tiago Luis, Fernandes Pablo, Mallmann Caroline Lorenci, Schons Cristine, Schuh Mateus Sabadi, Soares Pereira Rudiney, Fantinel Roberta Aparecida, Pereira da Silva Sally Deborah %J International Journal of Remote Sensing. Remotely Piloted Aircraft Systems (RPAS) and machine learning: A review in the context of forest science. 2021;42(21):8207-35.
17. Nanyonga Aziida, Wasswa Hassan, Turhan Ugur, Molloy Oleksandra, Wild Graham, editors. Sequential classification of aviation safety occurrences with natural language processing. AIAA AVIATION 2023 Forum; 2023.
18. Zhang Xiaoge, Mahadevan Sankaran %J Decision Support Systems. Ensemble machine learning models for aviation incident risk prediction. 2019;116:48-63.
19. Nanyonga Aziida, Wasswa Hassan, Molloy Oleksandra, Turhan Ugur, Wild Graham, editors. Natural language processing and deep learning models to classify phase of flight in aviation safety occurrences. 2023 IEEE Region 10 Symposium (TENSYMP); 2023: IEEE.
20. Nanyonga Aziida, Wild Graham %J arXiv preprint arXiv:.01222. Classification of Operational Records in Aviation Using Deep Learning Approaches. 2025.
21. Devlin Jacob %J arXiv preprint arXiv:.04805. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.
22. Iddrisu Abdul-Manan, Mensah Solomon, Boafo Fredrick, Yeluripati Govindha R, Kudjo Patrick %J International Journal of Information Management Data Insights. A sentiment analysis framework to classify instances of sarcastic sentiments within the aviation sector. 2023;3(2):100180.
23. Chandra Chetan, Ojima Yuga, Bendarkar Mayank V, Mavris Dimitri N %J Aerospace. Aviation-BERT-NER: Named Entity Recognition for Aviation Safety Reports. 2024;11(11):890.

 doi:10.20944/preprints202508.2196.v1

24. Qasim Rukhma, Bangyal Waqas Haider, Alqarni Mohammed A, Ali Almazroi Abdulwahab %J Journal of healthcare engineering. A fine-tuned BERT-based transfer learning approach for text classification. 2022;2022(1):3498123.
25. Kierszbaum Samuel, Lapasset Laurent, Klein Thierry %J Proc. CORIA. Transformer-based model on aviation incident reports. 2021.
26. González-Carvajal Santiago, Garrido-Merchán Eduardo C %J arXiv preprint arXiv:.13012. Comparing BERT against traditional machine learning text classification. 2020.
27. Rai Nishant, Kumar Deepika, Kaushik Naman, Raj Chandan, Ali Ahad %J International Journal of Cognitive Computing in Engineering. Fake News Classification using transformer based enhanced LSTM and BERT. 2022;3:98-105.
28. Kokab Sayyida Tabinda, Asghar Sohail, Naz Shehneela %J Array. Transformer-based deep learning models for the sentiment analysis of social media data. 2022;14:100157.
29. Oliaee Amir Hossein, Das Subasish, Liu Jinli, Rahman M Ashifur %J Natural language processing journal. Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types. 2023;3:100007.
30. Singh Utkarsha, Bhattacharya Margamitra, Padhi Radhakant. State-of-the-Art Natural Language Processing for Aviation: A Review.
31. Soyalp Gokhan, Alar Artun, Ozkanli Kaan, Yildiz Beytullah, editors. Improving text classification with transformer. 2021 6th International Conference on Computer Science and Engineering (UBMK); 2021: IEEE.
32. Gao Yubing, Zhu GuangYu, Duan Ya, Mao Jianfeng %J IEEE Transactions on Automation Science, Engineering. Semantic Encoding Algorithm for Classification and Retrieval of Aviation Safety Reports. 2024.
33. Liddy Elizabeth D. Natural language processing. 2001.
34. Nanyonga Aziida, Wasswa Hassan, Wild Graham, editors. Aviation Safety Enhancement via NLP & Deep Learning: Classifying Flight Phases in ATSB Safety Reports. 2023 Global Conference on Information Technologies and Communications (GCITC); 2023: IEEE.
35. Gupta Akhilesh, Tatbul Nesime, Marcus Ryan, Zhou Shengtian, Lee Insup, Gottschlich Justin. Class-weighted evaluation metrics for imbalanced data classification. 2020.
36. Kamyab Marjan, Liu Guohua, Adjeisah Michael %J Applied Sciences. Attention-based CNN and Bi-LSTM model based on TF-IDF and glove word embedding for sentiment analysis. 2021;11(23):11255.
37. Başarslan Muhammet Sinan, Kayaalp Fatih %J Journal of Cloud Computing. MBi-GRUMCONV: A novel Multi Bi-GRU and Multi CNN-Based deep learning model for social media sentiment analysis. 2023;12(1):5.
38. Loshchilov Ilya, Hutter Frank %J arXiv preprint arXiv:.05101. Fixing weight decay regularization in adam. 2017;5:5.
39. O'shea Keiron, Nash Ryan %J arXiv preprint arXiv:.08458. An introduction to convolutional neural networks. 2015.
40. Hochreiter S. J. Neural Computation M. I. T. Press. Long Short-term Memory. 1997.
41. Nanyonga Aziida, Wasswa Hassan, Joiner Keith, Turhan Ugur, Wild Graham %J Aerospace. Explainable Supervised Learning Models for Aviation Predictions in Australia. 2025;12(3):223.
42. Reza Selim, Ferreira Marta Campos, Machado José JM, Tavares João Manuel RS %J Expert Systems with Applications. A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. 2022;202:117275.
43. Vaswani A. J. Advances in Neural Information Processing Systems. Attention is all you need. 2017.
44. Rengasamy Divish, Morvan Hervé P, Figueredo Grazziela P, editors. Deep learning approaches to aircraft maintenance, repair and overhaul: A review. 2018 21st International Conference on Intelligent Transportation Systems (ITSC); 2018: IEEE.
45. Chawla Nitesh V, Bowyer Kevin W, Hall Lawrence O, Kegelmeyer W Philip %J Journal of artificial intelligence research. SMOTE: synthetic minority over-sampling technique. 2002;16:321-57.
46. Habbat Nassera, Nouri Hicham, Anoun Houda, Hassouni Larbi %J Engineering Applications of Artificial Intelligence. Sentiment analysis of imbalanced datasets using BERT and ensemble stacking for deep learning. 2023;126:106999.

 doi:10.20944/preprints202508.2196.v1

47. Ribeiro Marco Tulio, Singh Sameer, Guestrin Carlos, editors. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016.
48. Kannan Rithesh, Ng Hu, Yap Timothy Tzen Vun, Wong Lai Kuan, Chua Fang Fang, Goh Vik Tor, Lee Yee Lien, Wong Hwee Ling %J International Journal of Electrical, Engineering Computer. Handling class imbalance in education using data-level and deep learning methods. 2025;15(1):741-54.
49. Helmreich Robert L, Merritt Ashleigh C. Culture at work in aviation and medicine: National, organizational and professional influences: Routledge; 2017.
50. Somerville Alexander, Lynar Timothy, Wild Graham %J Transportation Engineering. The nature and costs of civil aviation flight training safety occurrences. 2023;12:100182.
51. Thai-Nghe Nguyen, Nghi DT, Schmidt-Thieme Lars, editors. Learning optimal threshold on resampling data to deal with class imbalance. Proc IEEE RIVF International Conference on Computing and Telecommunication Technologies; 2010.
52. Jeni László A, Cohn Jeffrey F, De La Torre Fernando, editors. Facing imbalanced data--recommendations for the use of performance metrics. 2013 Humaine association conference on affective computing and intelligent interaction; 2013: IEEE.
53. Vamvakas Panagiotis, Tsiropoulou Eirini Eleni, Papavassiliou Symeon %J Sensors. Risk-aware resource management in public safety networks. 2019;19(18):3853.