

## Article

# Natural Language Processing for Aviation Safety: Predicting Injury Levels from Incident Reports in Australia

Aziida Nanyonga<sup>1</sup>, Keith Joiner<sup>2</sup>, Ugur Turhan<sup>3</sup> and Graham Wild<sup>3,\*</sup>

<sup>1</sup> School of Engineering and Technology, University of New South Wales, Canberra, ACT 2600, Australia; a.nanyonga@unsw.edu.au

<sup>2</sup> Capability Systems Centre, University of New South Wales, Canberra, ACT 2610, Australia; k.joiner@unsw.edu.au

<sup>3</sup> School of Science, University of New South Wales, Canberra, ACT 2612, Australia; u.turhan@unsw.edu.au

\* Correspondence: g.wild@unsw.edu.au

**Abstract:** This study investigates the application of advanced deep learning models for the classification of aviation safety incidents, focusing on four models: Simple Recurrent Neural Network (sRNN), Gated Recurrent Unit (GRU), Bidirectional Long Short-Term Memory (BLSTM), and DistilBERT. The models were evaluated based on key performance metrics, including accuracy, precision, recall, and F1-score. DistilBERT achieved perfect performance with an accuracy of 1.00 across all metrics, while BLSTM demonstrated the highest performance among the deep learning models, with an accuracy of 0.9896, followed by GRU (0.9893) and sRNN (0.9887). Class-wise evaluations revealed that DistilBERT excelled across all injury categories, with BLSTM outperforming the other deep learning models, particularly in detecting fatal injuries, achieving a precision of 0.8684 and an F1-score of 0.7952. The study also addressed the challenges of class imbalance by applying class weighting, although the use of more sophisticated techniques, such as focal loss, is recommended for future work. This research highlights the potential of transformer-based models for aviation safety classification and provides a foundation for future research to improve model interpretability and generalizability across diverse datasets. These findings contribute to the growing body of research on applying deep learning techniques to aviation safety and underscore opportunities for further exploration.



Academic Editor: Alfredo Cuzzocrea

Received: 30 March 2025

Revised: 23 May 2025

Accepted: 26 May 2025

Published: 28 May 2025

**Citation:** Nanyonga, A.; Joiner, K.; Turhan, U.; Wild, G. Natural Language Processing for Aviation Safety: Predicting Injury Levels from Incident Reports in Australia. *Modelling* **2025**, *6*, 40. <https://doi.org/10.3390/modelling6020040>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** natural language processing; aviation safety; distilled BERT; sRNN

## 1. Introduction

Aviation safety is an area of significant concern due to the catastrophic consequences that accidents and incidents can have [1,2]. The repercussions of aviation accidents are severe, leading to fatalities, substantial financial losses, and long-term environmental damage [3]. Understanding the severity of safety occurrences is necessary for enhancing safety protocols, improving aircraft design, and informing regulatory frameworks [4]. Traditionally, aviation safety reports contain both structured and unstructured data. Structured data captures quantitative aspects such as aircraft type, location, and the phase of flight, while unstructured data, particularly textual narratives, provides qualitative insights into the context and causes of incidents [5]. However, extracting meaningful information from these unstructured descriptions remains a significant challenge, necessitating the development of automated methods for improving the timeliness of safety analysis, decision-making, and preventive actions [6].

A critical aspect of aviation safety lies in classifying injury levels sustained by individuals during safety occurrences. These range from minor injuries to fatal consequences, which serve as key indicators of incident severity [7]. Current classification methods primarily rely on manual or semi-automated techniques, which are labor-intensive, time-consuming, and susceptible to inconsistencies due to subjective interpretation [8]. However, with improvements in large-scale aviation safety datasets, there is an opportunity for more efficient, automated approaches to quickly and accurately infer injury levels from textual incident reports [9]. Natural language processing (NLP) and deep learning (DL) models offer scalable and consistent solutions for such textual analysis [10–12].

The primary objective of this study was to explore how advanced NLP and DL techniques can be employed to infer the injury levels sustained by individuals involved in aviation safety occurrences based solely on the textual narratives contained in incident reports. This research addresses the following question: To what extent can the injury levels be inferred from the textual narratives using state-of-the-art NLP techniques? NLP combined with DL architectures such as Simple Recurrent Neural Network (sRNN), Gated Recurrent Unit (GRU), or Bidirectional Long Short-Term Memory (BLSTM) likely offers a promising solution for automating the extraction of valuable information from these reports [13]. These models excel in learning from sequential data, enabling them to capture the intricate relationships between text and the severity of incidents, such as the injury levels sustained by individuals. Consequently, this study aims to demonstrate the feasibility and effectiveness of these advanced models in classifying injury levels from incident narratives. The unique challenge of inferring injury levels from textual narratives lies in the complexity and variability of language used in aviation safety reports. These documents frequently contain technical jargon, abbreviations, and context-dependent information [13].

The contributions of this research are twofold. (1) It demonstrates how DL models, such as DistilBERT, sRNN, BLSTM, and GRU, can be effectively employed to classify injury levels in aviation safety narratives. (2) It provides a scalable and automated solution to the problem of injury classification, which has traditionally relied on manual or semi-automated methods [14].

The remainder of the paper is structured as follows: Section 2 reviews the existing literature on NLP and DL in aviation safety, Section 3 details the dataset, preprocessing steps, and model architecture employed, Section 4 outlines the experimental results, Section 5 discusses the results and key findings, and Section 6 presents conclusions and outlines directions for future research. Beyond academia, the findings of this research offer practical implications for accident prevention, regulatory compliance, and the development of more effective safety measures in the aviation industry.

## 2. Related Work

NLP and ML techniques in aviation safety have garnered increasing attention in recent years, primarily to enhance safety analysis, decision-making, and risk management. Bloedorn [15] was a pioneer in applying ML to aviation safety reports, identifying patterns and predicting accident outcomes by extracting useful information from unstructured textual data. His work underscored the importance of leveraging NLP techniques for improving risk management. However, most studies since then have focused on broad classification tasks such as incident categorization based on general types or severity levels, rather than the specific challenge of predicting human injury levels (e.g., minor, serious, or fatal injuries). This research gap represents a key opportunity for advancing aviation safety research.

Nanyonga et al. [16] applied DL techniques to classify flight phases in safety occurrences. While their research demonstrated the feasibility of NLP techniques for safety data, it did not address the specific task of predicting injury levels, a key component of incident

severity and subsequent safety actions. Similarly, Zhang et al. [13] demonstrated the effectiveness of LSTM-based models for incident classification, highlighting the capability of DL models to manage complex aviation safety narratives, which often contain technical language and contextual nuances. However, the task of predicting injury levels based on these narratives remains underexplored.

Several studies have applied advanced DL methods, including RNNs, LSTMs, and GRUs, to aviation safety reports, achieving strong performance in incident classification and trend analysis [17–20]. These models excel in sequential text data analysis, extracting meaningful patterns from complex reports.

The potential of DistilBERT and other transformer-based models has also been explored in various domains. DistilBERT, an efficient version of BERT, has been applied to tasks like sentiment analysis and text classification [21], but not yet to injury level prediction in aviation safety.

Ahmad et al. [22] applied DL models to predict risk levels in transportation incidents using textual data, showcasing methodologies that share similarities with the task of injury level classification. This work, although not aviation-specific, demonstrates the promise of DL in classifying risk levels in safety-critical domains. The use of LSTMs and GRUs in sectors like healthcare and manufacturing has been shown to improve prediction accuracy in analyzing complex, sequential data [19,23].

A common challenge in applying NLP to aviation safety reports is the complexity of the text, which often involves technical terminology, abbreviations, and implicit contextual information. Traditional machine learning methods struggle with these complexities, but DL models, especially those that capture long-range dependencies in sequential data, have demonstrated superior performance [24]. Nanyonga et al. [25] used topic modeling to extract meaningful insights from aviation safety narratives, further emphasizing the utility of NLP in aviation.

While significant progress has been made in applying DL and NLP to aviation safety data, the task of predicting injury levels from textual descriptions remains an area with knowledge gaps. Studies have highlighted the promise of DL models like RNNs, LSTMs, GRUs, and DistilBERT in improving incident classification and safety analysis. This research aims to fill the gap, aiming to provide valuable insights for industry stakeholders and regulatory bodies to enhance the timeliness of risk assessment and prevention strategies.

### 3. Materials and Methods

To assess the efficacy of DL models in predicting injury levels from unstructured textual narratives in aviation safety reports, a comprehensive approach is employed, involving data collection, preprocessing, model selection, and performance evaluation as shown in Figure 1. The models considered in this study include traditional RNNs, LSTM networks, GRUs, and DistilBERT. Simpler and more traditional machine learning architectures, such as Naïve Bayes and Support Vector Machines (SVMs), were excluded as they are known to perform poorly in text classification tasks [26]. The implementation process, including performance evaluation, followed a systematic procedure to ensure robustness and reliability of the results.

#### 3.1. Data Collection

The dataset for this study is derived from the Australian Transport Safety Bureau (ATSB) database, consisting of over 53,275 aviation safety records. These records are provided in the form of incident reports and narratives, which describe the circumstances, causes, and outcomes of aviation safety occurrences. The dataset includes detailed textual descriptions of incidents, such as the aircraft type, location, phase of flight, and, importantly, the level of injury sustained by the individuals. Each safety occurrence is accompanied by a classification indicating the injury level, ranging from minor to fatal.

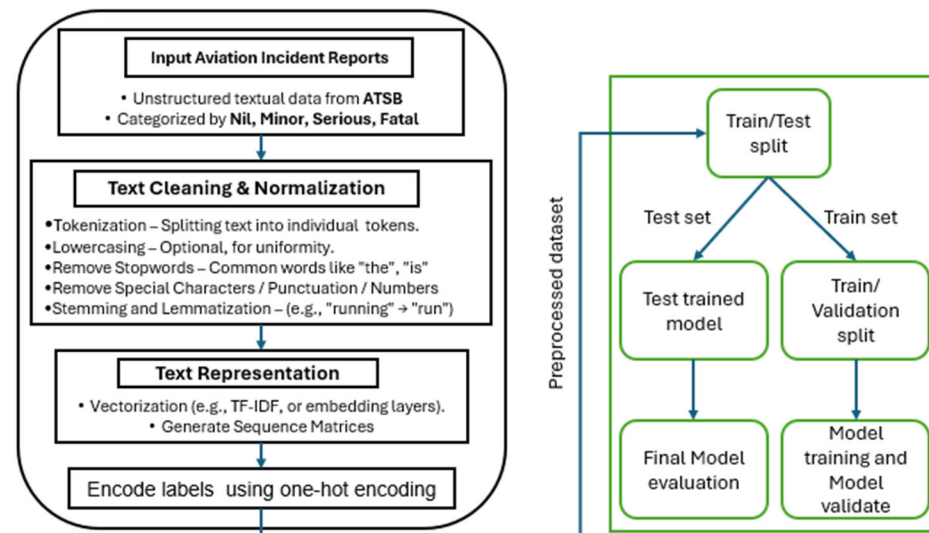


**Figure 1.** Methodological framework.

### 3.2. Text Preprocessing

Text preprocessing is a fundamental step in preparing unstructured text data for machine learning models. In this study, we leveraged the Keras deep learning (DL) library [27], as it provides a comprehensive suite of DL models and layers needed for text analysis. The Tokenizer module from Keras was employed to efficiently tokenize the input text and convert it into sequence vectors. This preprocessing step is vital for transforming raw textual data into a format suitable for machine learning algorithms.

To encode categorical data, such as the injury level labels (minor, serious, nil, and fatal), we used the *to\_categorical* module from Keras. This module maps categorical entries to numerical values using one-hot encoding, allowing the model to learn more efficiently by representing each label as a binary vector.

In addition to tokenization and encoding, we addressed challenges related to special characters, punctuation, and stop words using the *Spacy* library (version 3.8.0). *Spacy* is a powerful *Python* library tailored for text-processing tasks, including named entity recognition, word tagging, and lemmatization [28]. The library maintains an extensive list of stop words, punctuation marks, and special characters, with regular updates ensuring its continued relevance in processing modern text data. The lemmatization process helped reduce words to their base forms, aiding in generalizing the model.

Each input narrative underwent a comprehensive preprocessing pipeline to ensure that the text was uniformly represented. Specifically, the processed text was transformed into sequences or vectors with a fixed length of 2000 words [16]. Narratives with fewer than 2000 words were padded with zeros, while longer narratives were truncated to maintain uniformity. This consistent input length ensures that the neural network can process the data efficiently. The vocabulary size for the corpus was set to 10,000, allowing for the inclusion of a diverse set of terms and enhancing the model's ability to handle a broad range of aviation-related terminology. The dataset was split into training (80%) and testing (20%) sets to evaluate the model's performance [17]. This division ensured that the model was trained on a substantial portion of the data while allowing for unbiased evaluation on the excluded test set. Stratified sampling was applied during the split to preserve the distribution of the injury level classes (minor, serious, and fatal) across both sets.

To further ensure model robustness and mitigate overfitting during training, 10% of the training dataset was reserved for model validation in each epoch. This approach enabled continuous evaluation of the models, allowing for dynamic refinement and hyperparameter adjustment throughout the training process. The validation acted as a reliable

benchmark to detect overfitting and generalization gaps and then to refine the dropout rate per epoch [29]. Importantly, textual narratives were strictly assigned to only one set, either training, validation, or testing, ensuring that there was no overlap or data leakage between the sets. All performance metrics were computed exclusively on the unseen test set, which remained untouched throughout the model training and validation stages. This clear separation between training, validation, and testing data reinforces the reliability of the results [30]. Table 1 provides representative examples of aviation safety narratives and their corresponding injury levels. These records were selected based on the richness and length of their textual content to highlight the linguistic diversity and contextual depth present in the dataset.

**Table 1.** Sample records from the ATSB dataset.

| Ref.         | Date/Time            | Location                        | State | Phase of Flight | Summary                                                                                                                       | Injury Level | Departure → Destination               |
|--------------|----------------------|---------------------------------|-------|-----------------|-------------------------------------------------------------------------------------------------------------------------------|--------------|---------------------------------------|
| OA2013-00142 | 1/1/2013<br>12:01 AM | near Port Hedland Aerodrome     | WA    | Descent         | During the descent, the aircraft was struck by lightning.                                                                     | Nil          | Perth [YPPH] → Port Hedland [YPPD]    |
| OA2013-00167 | 1/1/2013<br>1:00 AM  | near Williamtown Aerodrome      | NSW   | Initial Climb   | During the initial climb, the aircraft encountered windshear.                                                                 | Unknown      | Williamtown [YWLM] → Melbourne [YMML] |
| OA2013-00196 | 1/1/2013<br>1:00 AM  | Bali International Airport      | Other | Standing        | The passenger declared undeclared fireworks on board.                                                                         | Nil          | Bali [WADD] → Perth [YPPH]            |
| OA2013-00045 | 1/5/2013<br>6:55 AM  | Darwin Aerodrome                | NT    | Initial Climb   | During the initial climb, the aircraft struck a bird.                                                                         | Nil          | Darwin [YPDN] → Dili [WPDL]           |
| OA2013-00055 | 1/5/2013<br>8:30 AM  | Ballina/Byron Gateway Aerodrome | NSW   | Approach        | During the final approach, the aircraft struck a swallow.                                                                     | Nil          | Sydney [YSSY]/ → Ballina/Byron [YBNA] |
| OA2013-00051 | 1/5/2013<br>9:30 AM  | Parafield Aerodrome             | SA    | Take-off        | During the take-off run, the aircraft struck a magpie.                                                                        | Nil          | Parafield [YPPF] → Parafield [YPPF]   |
| OA2013-00634 | 1/5/2013<br>10:15 AM | Perth Aerodrome                 | WA    | Initial Climb   | The crew received pitot static system warnings and returned. Engineering found no faults.                                     | Nil          | Perth [YPPH] → Sydney [YSSY]          |
| OA2013-00233 | 1/5/2013<br>10:40 AM | near Karratha Aerodrome         | WA    | Cruise          | The crew received a GPU warning during the cruise and returned. Inspection found that the GPU door was not latched correctly. | Nil          | Karratha [YPKA] → Karratha [YPKA]     |

### 3.3. Text Classification

In this study, four distinct DL architectures were researched for text classification tasks: GRU, BLSTM, sRNN, and DistilBERT. Each of these models offers unique advantages for processing and classifying text. GRU and BLSTM, both variants of RNNs, excel at capturing sequential dependencies and context in textual data [31], while sRNN provides a simpler, yet effective, architecture for time-series and sequence modeling [32]. DistilBERT, a transformer-based model, offers advanced performance by leveraging pretrained language representations and self-attention mechanisms, which are particularly well suited for handling long-range dependencies within text [21]. To address class imbalance in the dataset, class weights were applied across all models during training using the approach by [33]. This approach ensured that the models did not disproportionately favor majority classes, thereby improving the classification of underrepresented injury categories.

Model optimization was performed using the Adam optimizer, which was selected for its efficiency in gradient-based optimization and its ability to handle sparse gradients and adaptive learning rates as recommended by [34]. This optimizer, known for its speed and low memory requirements, was applied uniformly across all models. It is important to note that the focus of this study was not on optimizing the choice of the best optimizer, but rather on assessing the effectiveness of the chosen models for the text classification task [34]. Future research could explore alternative optimization techniques to further enhance model performance.

To enhance transparency and reproducibility, hyperparameter tuning was systematically conducted for each model. For the BLSTM model, experiments were performed with 1 to 3 layers, with 2 layers offering the most favorable performance, and a hidden unit size of 128 was selected after testing 64 and 256. A dropout rate of 0.3 yielded the best regularization effect from range testing from 0.2 to 0.5. The Adam optimizer was used with a learning rate of 0.0005. For the GRU model, the optimal configuration included 64 filters, a kernel size of 5, max pooling, a dropout rate of 0.3, and a learning rate of 0.001 with a batch size of 32, and the sRNN followed similar tuning guidelines. For the DistilBERT model, fine-tuning involved testing batch sizes of 8, 16, and 32, with 16 proving optimal, and learning rates of  $3 \times 10^{-5}$ ,  $5 \times 10^{-5}$ , and  $1 \times 10^{-4}$ , where  $2 \times 10^{-5}$  delivered the best performance. The DistilBERT model was trained in over 3 epochs with a maximum sequence length of 256. These settings ensured that each model was trained under conditions optimized for both accuracy and efficiency.

### 3.4. Deep Learning Architecture

To ensure consistency and comparison across all models, a unified DL architecture was employed as the baseline, with minor modifications made to accommodate the specific requirements of each model in Figure 2. This standardized architecture comprised three fundamental components: an embedding layer, multiple hidden layers, and an output layer.

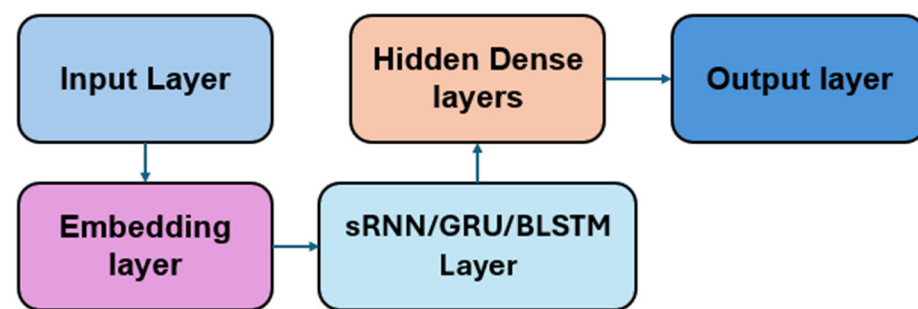


Figure 2. DL model architecture.

The embedding layer was utilized to convert the input text sequences into dense vectors of fixed dimensionality, enabling the model to capture both semantic and syntactic relationships. After the embedding layer, the hidden layers incorporated the Rectified Linear Unit (ReLU) activation function. ReLU was selected for its ability to introduce non-linearity, thereby facilitating the model's capacity to learn intricate patterns and complex dependencies within the data [35]. Additionally, ReLU mitigates issues associated with vanishing gradients, promoting efficient training and faster convergence.

The output layer employed the *SoftMax* (version 7.3.) activation function as it is particularly well suited for multi-class classification tasks [36]. This function produces a probability distribution over the possible output classes, with the highest probability corresponding to the model's predicted class. The final classification decision was made by applying the *argmax* function, which identifies the index corresponding to the maximum probability within the *SoftMax* output, thereby selecting the most probable class for each input sample.

#### 3.4.1. RNN

The first DL architecture considered was the standard RNN, a fundamental type of RNN architecture that processes sequential data by feeding the output of the previous time step as input to the current time step. Its architecture is relatively basic, consisting of a single hidden layer that facilitates the flow of information from one step to the next [37]. Despite their advantages in handling sequential data, standard RNNs suffer from limitations in capturing long-range dependencies due to the vanishing gradient problem [38].

### 3.4.2. BLSTM

A BLSTM network evolved to overcome the shortcomings of the standard RNN. LSTM is a specialized form of RNN and is designed to mitigate the vanishing gradient problem and capture long-range dependencies in sequential data [11]. BLSTM architecture extends the standard LSTM by processing the input sequence in both forward and backward directions, allowing the model to assign weight to contexts from both the past and future [39].

### 3.4.3. GRU

Another architecture considered is the GRU, a variant of the LSTM that simplifies the model by combining the forget and input gates into a single update gate. GRUs, therefore, have fewer parameters than LSTMs, making them computationally more efficient while retaining the ability to capture long-range dependencies in the input data [40].

### 3.4.4. DistilBERT

To explore cutting-edge NLP models, this study also incorporates DistilBERT, a smaller, faster, and more efficient version of the original BERT model. BERT (Bidirectional Encoder Representations from Transformers) revolutionized NLP by introducing a deep bidirectional transformer architecture that captures contextual relationships from both the left and right of a token's position within a text. Due to this bidirectional capability and extensive pretraining on large corpora, BERT has set new benchmarks across a variety of NLP tasks [41]. DistilBERT is a distilled version of BERT, developed through a knowledge distillation process where a smaller model (student) learns to replicate the behavior of a larger, pretrained model (teacher). Although it has 40% fewer parameters and is 60% faster than BERT, DistilBERT retains about 97% of BERT's language understanding capabilities [21]. To provide clarity on the underlying mechanism, Figure 3 illustrates the core architecture of the transformer encoder that underpins both BERT and DistilBERT. The transformer encoder begins with an input embedding layer, which incorporates word embeddings, positional embeddings (to capture the order of tokens), and optionally, type embeddings (though DistilBERT omits type embeddings to reduce complexity). These embeddings are then passed into a stack of transformer blocks. Each transformer block comprises two primary components: a multi-head self-attention mechanism and a position-wise feed-forward neural network. The multi-head attention mechanism allows the model to focus on different parts of the input sequence simultaneously, capturing various aspects of contextual relationships between words. This is followed by a feed-forward network applied to each position independently. Both sub-layers are surrounded by residual connections and layer normalization (denoted as "Add & Norm" in the figure) to facilitate stable and efficient training.

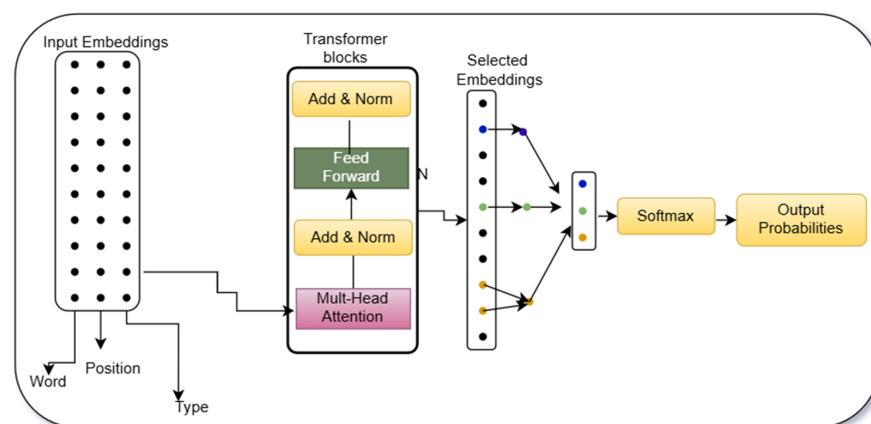


**Figure 3.** BERT model architecture.

Finally, the output of the last transformer layer is passed through a classification head, which includes a *Softmax* layer to produce output probabilities corresponding to different injury level classes. This design allows the model to make predictions based on a nuanced understanding of the input narratives, while remaining efficient and scalable due to its distilled form.

### 3.5. Model Implementation

This study was implemented using Python (version 3.8.10), leveraging a variety of machine learning and DL libraries for data preprocessing, model training, evaluation, and visualization. DL models were trained using *TensorFlow* (version 2.10.0) and *Keras* (version 2.10.0), while the BERT model was fine-tuned using 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. Visualizations of the model performance were generated with *Matplotlib* (version 3.6.1) and *Seaborn* (version 0.12.0). 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) was employed. The experiments were conducted within a *Jupyter* Notebook environment (version 7.4.3), which was hosted on a *Linux* server equipped with 256 CPU cores, 256 GB of RAM, and running *Ubuntu* (version 5.4.0-169-generic).

### 3.6. Model Performance Evaluation

To ensure a rigorous assessment of the models' effectiveness, the well-established performance metrics of accuracy, precision, recall, and F1-score were used in Table 2. Accuracy is the proportion of correct predictions (both true positives and true negatives) relative to the total number of predictions. Precision, which focuses on the quality of positive predictions, is defined as the proportion of true positives among all instances predicted as positive (i.e., true positives and false positives). Recall, or sensitivity, evaluates the model's ability to identify positive instances, measuring the proportion of true positives among all actual positive instances (i.e., true positives and false negatives). The F1-score balances precision and recall, which is particularly useful in cases of class imbalance, as it helps prevent favoring the majority class.

**Table 2.** Performance evaluation metrics.

| Metrics               | Evaluation Focus                                   | Formula                                                                         |
|-----------------------|----------------------------------------------------|---------------------------------------------------------------------------------|
| <b>Precision (p)</b>  | Correctly predicted positives in a positive class  | $\frac{TP}{TP+FP}$                                                              |
| <b>Recall (r)</b>     | Fraction of positive patterns correctly classified | $\frac{TP}{TP+FN}$                                                              |
| <b>F1-score (F)</b>   | Weighted average score of precision and recall     | $\frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$ |
| <b>Accuracy (acc)</b> | Total number of instances predicted correctly      | $\frac{TP+TN}{TP+FP+TN+FN}$                                                     |

These metrics allow for a thorough analysis of each model's classification performance, not only in terms of overall correctness but also in terms of the model's ability to correctly classify each injury level.

Beyond these standard performance metrics, we further evaluated the classification results by analyzing confusion matrices to explore the distribution of prediction errors, particularly false positives and false negatives, across injury categories. In the context of aviation safety, the impact of these errors can be significant. A false negative, where a serious or fatal incident is misclassified as *minor* or *nil*, poses a higher operational

risk compared to a false positive, which may only lead to overestimating an incident's severity. Therefore, in real-world safety-critical applications, minimizing false negatives becomes a crucial objective. We have incorporated this consideration into our model evaluation strategy and tuning process, particularly for the high-risk classes of *serious* and *fatal*. This targeted focus ensures that the models are accurate overall and effective at identifying critical safety events, which is vital for timely intervention and risk mitigation in aviation operations.

## 4. Results

This section presents a comprehensive evaluation of the models' performance in classifying aviation safety incidents. The analysis is structured to first provide an overview of the models' overall classification performance, followed by a class-wise assessment of their predictive capabilities. Also, validation accuracy and loss trends are examined to evaluate the models' convergence and generalization.

### 4.1. Overall Model Performance

Table 3 presents a comparative evaluation of the four models, sRNN, GRU, BLSTM, and DistilBERT, based on the main performance metrics. The results indicate that DistilBERT outperforms all other models, achieving perfect scores (1.00) across all metrics, thereby demonstrating its exceptional capability in accurately classifying aviation safety incidents. If there is overfitting in this model, our methods could not detect it, and potentially more enhanced validation methods need to be researched using greater computation than was available to this study.

**Table 3.** Overall model performance.

| Model      | Accuracy    | Precision   | Recall      | F1-Score    |
|------------|-------------|-------------|-------------|-------------|
| sRNN       | 0.9887      | 0.9886      | 0.9887      | 0.9885      |
| GRU        | 0.9893      | 0.9891      | 0.9893      | 0.9891      |
| BLSTM      | 0.9896      | 0.9893      | 0.9896      | 0.9894      |
| DistilBERT | <b>1.00</b> | <b>1.00</b> | <b>1.00</b> | <b>1.00</b> |

Among the DL models, BLSTM exhibited the highest classification accuracy at 0.9896, followed closely by GRU (0.9893) and sRNN (0.9887). The minimal variations in performance between these models suggest that all of them provide robust classification capabilities, though BLSTM demonstrates slightly better generalization. These findings underscore the effectiveness of DL models in handling sequential data, with BLSTM benefiting from its bidirectional processing ability, allowing it to capture contextual dependencies more effectively.

### 4.2. Class-Wise Performance Evaluation

A more detailed evaluation of model performance across individual injury severity categories (nil, minor, fatal, and serious) is provided in Table 4. DistilBERT again demonstrates flawless classification for all injury levels. This result confirms its superior ability to capture nuanced patterns in aviation safety incident narratives.

Among the DL models, BLSTM consistently outperformed GRU and sRNN, particularly in detecting the *fatal* and *serious* injury categories. The GRU model exhibited the highest precision in classifying *fatal* injuries (0.9167), exceeding that of sRNN (0.9062). However, in terms of recall for *minor* injuries, BLSTM achieved the highest score (0.7178), signifying its ability to correctly identify more instances of this category.

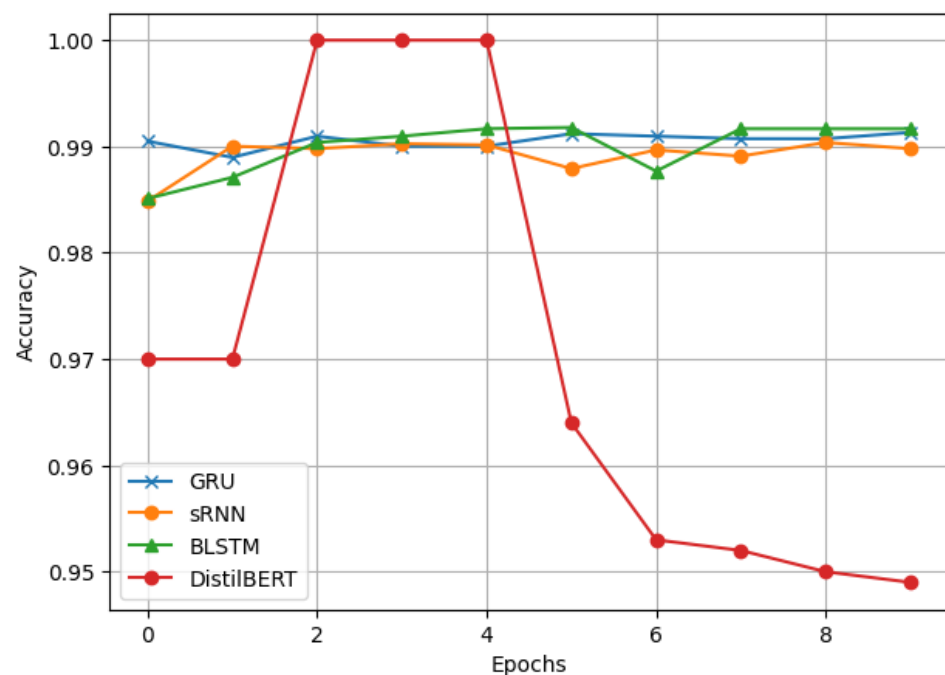
**Table 4.** Class-wise model performance.

| Model      | Metric    | Nil    | Minor  | Fatal  | Serious |
|------------|-----------|--------|--------|--------|---------|
| BLSTM      | Precision | 0.9944 | 0.7548 | 0.8684 | 0.7917  |
|            | Recall    | 0.9959 | 0.7178 | 0.7333 | 0.7917  |
|            | F1-Score  | 0.9951 | 0.7358 | 0.7952 | 0.7917  |
| sRNN       | Precision | 0.9942 | 0.7273 | 0.9062 | 0.7222  |
|            | Recall    | 0.9958 | 0.6871 | 0.6444 | 0.8125  |
|            | F1-Score  | 0.9950 | 0.7066 | 0.7532 | 0.7647  |
| GRU        | Precision | 0.9942 | 0.7197 | 0.9167 | 0.8478  |
|            | Recall    | 0.9959 | 0.6933 | 0.7333 | 0.8125  |
|            | F1-Score  | 0.9951 | 0.7063 | 0.8148 | 0.8298  |
| DistilBERT | Precision | 1.00   | 1.00   | 1.00   | 1.00    |
|            | Recall    | 1.00   | 1.00   | 1.00   | 1.00    |
|            | F1-Score  | 1.00   | 1.00   | 1.00   | 1.00    |

The results further highlight challenges associated with classifying less frequent categories, such as *fatal* and *serious* injuries. While precision scores remain relatively high across all models, the recall values for these categories indicate room for improvement, suggesting a need for further optimization, such as class-balancing techniques or advanced feature engineering, to mitigate potential biases in model learning.

#### 4.3. Model Validation Performance

To further demonstrate the robustness of the models, Figures 4 and 5 demonstrate the validation accuracy and loss trends across training epochs. DistilBERT achieved perfect accuracy during epochs 2 through 5, with epoch 3 being the most optimal, as the model maintained peak performance during this epoch. However, after epoch 5, a decline in accuracy was observed, continuing through to epoch 10, suggesting that overfitting may have occurred beyond this point. The high performance may be partly due to strong lexical patterns in the ATSB reports, which could be well suited to the DistilBERT model's capabilities, whereby further research is warranted on how this has occurred.

**Figure 4.** Validation accuracy of each model.

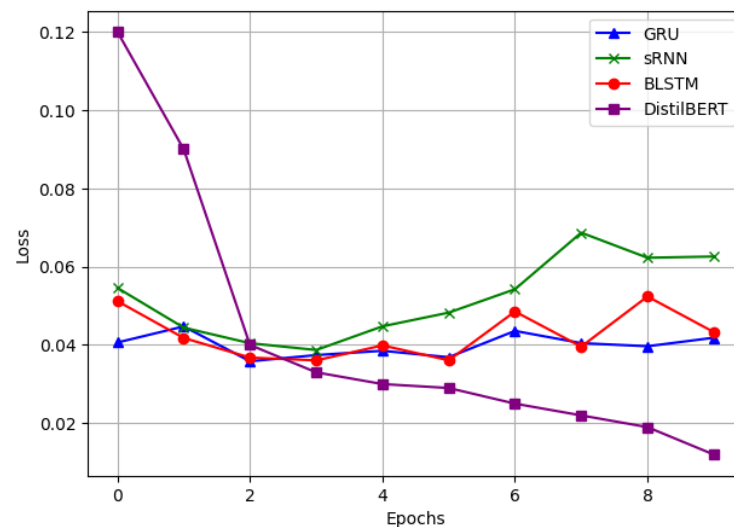


**Figure 5.** Validation loss of each model.

In contrast, the DL models exhibited gradual improvements in accuracy over successive epochs. The BLSTM model emerged as the highest-performing model among them, maintaining a steady increase in accuracy and showing superior generalization capability compared to the other models. This finding highlights the varying training dynamics and performance characteristics of transformer-based models like DistilBERT when compared to more traditional DL architectures like BLSTM and sRNN.

DistilBERT demonstrates a sharp reduction in validation loss early in training, reaching near-zero values as shown in Figure 5. This result indicates its ability to effectively capture linguistic structures with minimal overfitting. Among DL models, BLSTM exhibits the lowest validation loss, signifying its superior convergence and improved stability compared to GRU and sRNN. These findings reinforce the suitability of BLSTMs for handling aviation safety data, as they maintain high accuracy while effectively minimizing classification errors.

#### 4.4. Comparative Analysis of Model Performance

To further assess the classification capabilities of the DL models, Figure 6 presents the confusion matrices for BLSTM, the highest-performing DL model. This matrix provides an in-depth examination of the model's classification tendencies across injury categories.

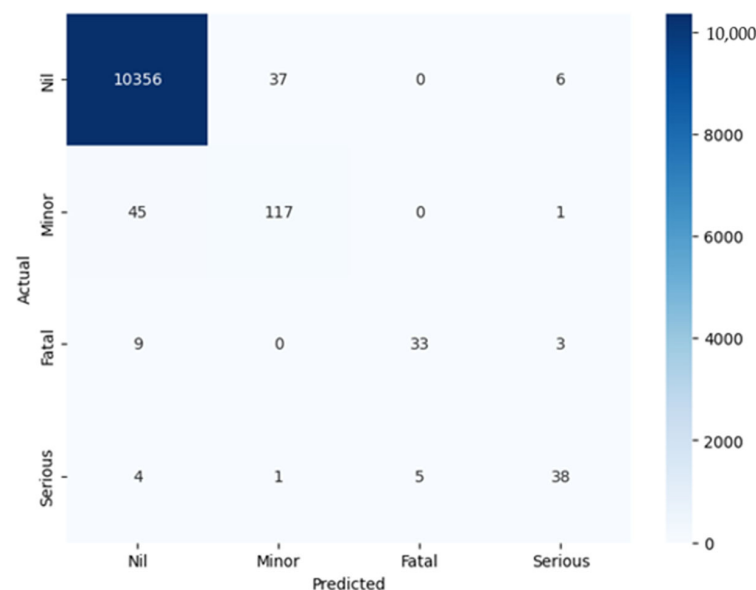


**Figure 6.** Confusion matrix for BLSTM.

The results reveal that BLSTM performs exceptionally well in classifying *nil* and *minor* injury categories, with minimal misclassifications. However, some degree of misclassification is observed for *fatal* and *serious* injuries. This deterioration can likely be attributed to class imbalances within the dataset, where instances of *fatal* and *serious* injuries are relatively underrepresented. Despite these challenges, BLSTM maintains a high level of predictive accuracy, suggesting that further refinement, such as data augmentation or weighted loss functions [42], could enhance its classification performance in underrepresented categories.

## 5. Ablation

To better understand the contributions of various modeling approaches, we conducted an ablation study evaluating the impact of different DL architectures and techniques on classification performance [43]. One key aspect examined was the effect of class weighting when applied across all models to mitigate the impact of class imbalance. While class weighting improved recall for underrepresented classes, particularly for the *fatal* and *serious* injury categories, some degree of misclassification persisted, suggesting that more sophisticated imbalance-handling techniques, such as focal loss [44] or data augmentation methods [45], could further enhance model robustness.

Additionally, the study compared traditional recurrent models such as sRNN, GRU, and BLSTM against DistilBERT, a transformer-based model. The results demonstrated that while DL architectures like BLSTM exhibited strong classification performance, they remained susceptible to minor variations in minority class predictions. DistilBERT, on the other hand, achieved perfect classification accuracy, likely due to its ability to capture contextual dependencies more effectively through self-attention mechanisms [46].

We further examined the impact of pretraining on model performance. The superior results of DistilBERT highlight the importance of leveraging pretrained language models for aviation safety text classification, as they provide a richer semantic understanding of domain-specific narratives [47]. In contrast, recurrent models trained from scratch required significantly more computational effort to reach near-optimal performance. These findings align with previous studies emphasizing the advantages of transformer-based architectures in NLP tasks [48].

### 5.1. Discussion

The findings of this study demonstrate the general efficacy of advanced DL models for aviation safety text classification. The fact that DistilBERT significantly outperforms traditional recurrent architectures aligns with prior research highlighting the advantages of transformer-based models in text classification tasks [49,50]. The ability of DistilBERT to maintain high precision, recall, and F1-scores suggests that self-attention mechanisms effectively capture long-range dependencies in the long and complex textual descriptions of aviation incidents, reducing ambiguity in classification.

When compared to sRNN and GRU models, particularly in classifying *fatal* and *serious* injury levels, the bidirectional structure of BLSTM allowed for more effective feature extraction, capturing both past and future contextual information. This is consistent with previous studies where BLSTM has been shown to outperform unidirectional recurrent models in sequential text processing [11,51]. However, despite the improved performance, BLSTM and the other DL models exhibited slight inconsistencies in minority class classification, emphasizing the challenges posed by imbalanced datasets in aviation safety analysis.

Model validation performance further reinforced these findings. The validation loss and accuracy curves revealed that DistilBERT achieved rapid convergence, indicating its strong generalization ability, while DL models such as BLSTM struggled to stabilize. This suggests that while DL models effectively capture sequential patterns, they may struggle

with distinguishing rare but critical cases. This observation warrants further investigation into specialized strategies for handling imbalance.

However, a noted limitation of this study is the lack of comparison with other recent and powerful transformer-based or large language models (LLMs) beyond DistilBERT. While DistilBERT's performance is impressive, the absence of models such as BERT [48], RoBERTa [52], DeBERTa [53], or GPT-style architectures [54] leaves open the question of whether DistilBERT truly represents the current state of the art. Including these benchmarks would have strengthened the study's claims and contextualized DistilBERT's performance more comprehensively. The decision to exclude larger models was primarily based on their higher computational costs and tuning complexity, which may limit their practicality in real-time aviation applications. Nevertheless, future research should incorporate these recent transformer variants to evaluate the scalability, robustness, and true comparative value of DistilBERT in aviation NLP tasks.

### 5.2. Limitations

As shown in the ablation study, DistilBERT achieved perfect accuracy; the generalizability of such transformer-based models in real-world applications needs to be validated on external datasets. Given that aviation safety reports vary across regions and regulatory bodies, future work should investigate cross-domain generalization.

Another limitation concerns computational efficiency. While DistilBERT exhibited superior performance, its deployment in real-time aviation safety monitoring systems requires careful consideration of computational costs. Transformer-based models are known to be resource-intensive, and their real-world application may necessitate optimization techniques such as model distillation or pruning [21]. Further research is needed to explore efficient implementations of these models for large-scale aviation safety monitoring.

Last, the study relied on textual data from the ATSB dataset, which, while comprehensive, may not capture the full spectrum of aviation safety incidents globally. Future research should incorporate multi-source datasets, including reports from different aviation authorities, to develop more universally applicable models. Additionally, explainability methods should be explored to enhance trust and interpretability in AI-driven aviation safety classification systems [55].

## 6. Conclusions and Future Work

The results of this study reinforce the effectiveness of advanced DL architectures in aviation safety text classification, this time in classifying injury prediction. The comparison of traditional recurrent neural networks and transformer-based models highlights the superior performance of DistilBERT, achieving perfect classification accuracy across all injury severity categories. The bidirectional structure of BLSTM also exhibited strong performance, outperforming the simpler recurrent sRNN and GRU models. These findings confirm the benefits of self-attention mechanisms and pretrained language models in handling textual data from aviation safety reports.

Despite these advancements, the study has several limitations. While class weighting was implemented to mitigate class imbalance, more sophisticated techniques such as focal loss or synthetic oversampling could further enhance model robustness. Additionally, the high computational demands of transformer-based models present challenges for real-time applications, necessitating further research into model optimization techniques. Moreover, the dataset used in this study, derived from the ATSB database, may not generalize across different regulatory bodies or international contexts. Future studies should explore multi-source datasets and domain adaptation techniques to improve generalizability.

Future work should also focus on enhancing the explainability of AI-driven aviation safety models. The application of interpretable machine learning methods, such as SHAP (Shapley Additive Explanations) [55] and LIME (Local Interpretable Model-Agnostic Explanations) [56], will be explored to increase transparency and stakeholder trust, particularly in high-risk categories like *serious* and *fatal*. Incorporating such tools can facilitate regulatory acceptance by offering insights into how specific tokens or phrases in incident narratives influence model predictions. Additionally, integrating multimodal data, such as sensor readings, flight parameters, and environmental conditions, could further refine predictive capabilities. Expanding the application of transformer-based architectures to broader aviation safety tasks, including risk assessment and predictive maintenance, presents an exciting avenue for future research. There is also potential for more advanced techniques, as covered by [57], such as SMOTE, oversampling, undersampling, and focal loss, which were not applied in this study but will be explored in future research.

**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; 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:

|            |                                                                   |
|------------|-------------------------------------------------------------------|
| BLSTM      | Bidirectional Long Short-Term Memory                              |
| DistilBERT | Distilled Bidirectional Encoder Representations from Transformers |
| FP         | False Positive                                                    |
| FN         | False Negative                                                    |
| GRU        | Gated Recurrent Unit                                              |
| LSTM       | Long Short-Term Memory                                            |
| ML         | Machine Learning                                                  |
| NLP        | Natural Language Processing                                       |
| ReLU       | Rectified Linear Unit                                             |
| sRNN       | Simple Recurrent Neural Network                                   |
| TP         | True Positive                                                     |
| TN         | True Negative                                                     |

## References

1. Čokorilo, O.; Gvozdenović, S.; Vasov, L.; Mirosavljević, P. Costs of unsafety in aviation. *Technol. Econ. Dev. Econ.* **2010**, *16*, 188–201. [CrossRef]
2. Somerville, A.; Lynar, T.; Wild, G. The nature and costs of civil aviation flight training safety occurrences. *Transp. Eng.* **2023**, *12*, 100182. [CrossRef]
3. Harris, D.; Li, W.-C. Using Neural Networks to predict HFACS unsafe acts from the pre-conditions of unsafe acts. *Ergonomics* **2019**, *62*, 181–191. [CrossRef] [PubMed]
4. Shappell, S.; Detwiler, C.; Holcomb, K.; Hackworth, C.; Boquet, A.; Wiegmann, D.A. Human error and commercial aviation accidents: An analysis using the human factors analysis and classification system. In *Human Error in Aviation*; Routledge: London, UK, 2017; pp. 73–88.

5. Nanyonga, A.; Joiner, K.; Turhan, U.; Wild, G. Applications of natural language processing in aviation safety: A review and qualitative analysis. In Proceedings of the AIAA SCITECH 2025 Forum, Orlando, FL, USA, 6–10 January 2025; p. 2153.
6. Xiong, M.; Wang, H.; Wong, Y.D.; Hou, Z. Enhancing aviation safety and mitigating accidents: A study on aviation safety hazard identification. *Adv. Eng. Inform.* **2024**, *62*, 102732. [[CrossRef](#)]
7. Slikboer, R.; Muir, S.D.; Silva, S.S.M.; Meyer, D. A systematic review of statistical models and outcomes of predicting fatal and serious injury crashes from driver crash and offense history data. *Syst. Rev.* **2020**, *9*, 220. [[CrossRef](#)]
8. Kazi, N.M.S. Using Machine Learning Models to Study Human Error Related Factors in Aviation Accidents and Incidents. Doctoral Dissertation, National College of Ireland, Dublin, Ireland, 2020.
9. Zhang, C.; Liu, C.; Liu, H.; Jiang, C.; Fu, L.; Wen, C.; Cao, W. Incorporation of pilot factors into risk analysis of civil aviation accidents from 2008 to 2020: A data-driven Bayesian network approach. *Aerospace* **2022**, *10*, 9. [[CrossRef](#)]
10. Nanyonga, A.; Wasswa, H.; Joiner, K.; Turhan, U.; Wild, G. A Multi-Head Attention-Based Transformer Model for Predicting Causes in Aviation Incident. *Modelling* **2025**, *6*, 27. [[CrossRef](#)]
11. Hochreiter, S.; Schmidhuber, J. Long Short-term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
12. Nanyonga, A.; Wasswa, H.; Turhan, U.; Joiner, K.; Wild, G. Exploring Aviation Incident Narratives Using Topic Modeling and Clustering Techniques. In Proceedings of the 2024 IEEE Region 10 Symposium (TENSYP), New Delhi, India, 27–29 September 2024; pp. 1–6.
13. Zhang, X.; Srinivasan, P.; Mahadevan, S. Sequential deep learning from NTSB reports for aviation safety prognosis. *Saf. Sci.* **2021**, *142*, 105390. [[CrossRef](#)]
14. Paul, S.; Purkaystha, B.S.; Das, P. NLP TOOLS USED IN CIVIL AVIATION: A SURVEY. *Int. J. Adv. Res. Comput. Sci.* **2018**, *9*, 109–114. [[CrossRef](#)]
15. Bloedorn, E. Mining aviation safety data: A hybrid approach. In Proceedings of the Armed Forces Communications and Electronics Association (AFCEA) First Federal Data Mining Symposium, Washington, DC, USA, 28–29 March 2000.
16. Nanyonga, A.; Wasswa, H.; Wild, G. Phase of Flight Classification in Aviation Safety Using LSTM, GRU, and BiLSTM: A Case Study with ASN Dataset. In Proceedings of the 2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS), Macau, China, 6–8 December 2023; pp. 24–28.
17. Nanyonga, A.; Wild, G. Classification of Operational Records in Aviation Using Deep Learning Approaches. *arXiv* **2025**, arXiv:2501.01222.
18. Kannan, R.; Ng, H.; Yap, T.T.V.; Wong, L.K.; Chua, F.F.; Goh, V.T.; Lee, Y.L.; Wong, H.L. Handling class imbalance in education using data-level and deep learning methods. *Int. J. Electr. Comput. Eng. (IJECE)* **2025**, *15*, 741–754. [[CrossRef](#)]
19. Ali, A.R.; Kamal, H. Time-to-Fault Prediction Framework for Automated Manufacturing in Humanoid Robotics Using Deep Learning. *Technologies* **2025**, *13*, 42. [[CrossRef](#)]
20. Nanyonga, A.; Wild, G. Impact of Dataset Size & Data Source on Aviation Safety Incident Prediction Models with Natural Language Processing. In Proceedings of the 2023 Global Conference on Information Technologies and Communications (GCITC), Bengaluru, India, 1–3 December 2023; pp. 1–7.
21. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
22. Ahmad, I.; Alqurashi, F.; Abozinadah, E.; Mehmood, R. Deep journalism and DeepJournal V1. 0: A data-driven deep learning approach to discover parameters for transportation. *Sustainability* **2022**, *14*, 5711. [[CrossRef](#)]
23. Rehman, A.; Saba, T.; Mujahid, M.; Alamri, F.S.; ElHakim, N. Parkinson’s disease detection using hybrid LSTM-GRU deep learning model. *Electronics* **2023**, *12*, 2856. [[CrossRef](#)]
24. Zhong, B.; Pan, X.; Love, P.E.; Sun, J.; Tao, C. Hazard analysis: A deep learning and text mining framework for accident prevention. *Adv. Eng. Inform.* **2020**, *46*, 101152. [[CrossRef](#)]
25. Nanyonga, A.; Wild, G. Analyzing Aviation Safety Narratives with LDA, NMF and PLSA: A Case Study Using Socrata Datasets. *arXiv* **2025**, arXiv:2501.01690.
26. Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.S.; He, L. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.* **2022**, *13*, 1–41. [[CrossRef](#)]
27. Jin, H.; Chollet, F.; Song, Q.; Hu, X. Autokeras: An automl library for deep learning. *J. Mach. Learn. Res.* **2023**, *24*, 1–6.
28. Hardeniya, N.; Perkins, J.; Chopra, D.; Joshi, N.; Mathur, I. *Natural Language Processing: Python and NLTK*; Packt Publishing Ltd.: Birmingham, UK, 2016.
29. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
30. Géron, A. Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts. 2019. Available online: <https://books.google.com.tr/books> (accessed on 20 June 2023).
31. Huang, C.-G.; Huang, H.-Z.; Li, Y.-F. A bidirectional LSTM prognostics method under multiple operational conditions. *IEEE Trans. Ind. Electron.* **2019**, *66*, 8792–8802. [[CrossRef](#)]

32. Fang, W.; Chen, Y.; Xue, Q. Survey on research of RNN-based spatio-temporal sequence prediction algorithms. *J. Big Data* **2021**, *3*, 97. [\[CrossRef\]](#)
33. Gupta, A.; Tatbul, N.; Marcus, R.; Zhou, S.; Lee, I.; Gottschlich, J. Class-weighted evaluation metrics for imbalanced data classification. In Proceedings of the ICLR 2021 Conference, Vienna, Austria, 4 May 2021.
34. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
35. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
36. Asadi, B.; Jiang, H. On approximation capabilities of ReLU activation and softmax output layer in neural networks. *arXiv* **2020**, arXiv:2002.04060.
37. Salem, F.M. A basic recurrent neural network model. *arXiv* **2016**, arXiv:1612.09022.
38. Socher, R.; Bengio, Y.; Manning, C.D. Deep learning for NLP (without magic). In Proceedings of the Tutorial Abstracts of ACL 2012, Jeju Island, Republic of Korea, 8 July 2012; p. 5.
39. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [\[CrossRef\]](#)
40. Dey, R.; Salem, F.M. Gate-variants of gated recurrent unit (GRU) neural networks. In Proceedings of the 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Boston, MA, USA, 6–9 August 2017; pp. 1597–1600.
41. Qasim, R.; Bangyal, W.H.; Alqarni, M.A.; Ali Almazroi, A. A fine-tuned BERT-based transfer learning approach for text classification. *J. Healthc. Eng.* **2022**, *2022*, 3498123. [\[CrossRef\]](#)
42. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Eng. Data* **2009**, *21*, 1263–1284.
43. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should i trust you?” Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
44. Ross, T.-Y.; Dollár, G. Focal loss for dense object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2980–2988.
45. Zhu, X.; Men, J.; Yang, L.; Li, K. Imbalanced driving scene recognition with class focal loss and data augmentation. *Int. J. Mach. Learn. Cybern.* **2022**, *13*, 2957–2975. [\[CrossRef\]](#)
46. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
47. Khandelwal, U.; Clark, K.; Jurafsky, D.; Kaiser, L. Sample efficient text summarization using a single pre-trained transformer. *arXiv* **2019**, arXiv:1905.08836.
48. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
49. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. *Adv. Neural Inf. Process. Systems* **2020**, *33*, 1877–1901.
50. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
51. Graves, A.; Mohamed, A.-r.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, USA, 26–31 May 2013; pp. 6645–6649.
52. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
53. He, P.; Liu, X.; Gao, J.; Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv* **2020**, arXiv:2006.03654.
54. Szentmihályi, K.; Nemeskey, D.M.; Szekeres, A.M.; Gortka, B.G.; Indig, B.; Palkó, G.; Nagy, B.; Vidács, L. Pretraining GPT-style models in Hungarian. *Infocommun. J.* **2025**, *17*, 2–11.
55. Nanyonga, A.; Wasswa, H.; Joiner, K.; Turhan, U.; Wild, G. Explainable Supervised Learning Models for Aviation Predictions in Australia. *Aerospace* **2025**, *12*, 223. [\[CrossRef\]](#)
56. Salih, A.M.; Raisi-Estabragh, Z.; Galazzo, I.B.; Radeva, P.; Petersen, S.E.; Lekadir, K.; Menegaz, G. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Adv. Intell. Syst.* **2025**, *7*, 2400304. [\[CrossRef\]](#)
57. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [\[CrossRef\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.