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

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

Posted Date: 31 March 2025

doi: 10.20944/preprints202503.2251.v1

Keywords: Natural Language Processing; Aviation Safety; Distilled BERT; sRNN

Article

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

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: 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 in NLP-driven risk assessment and incident classification.

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 on human life, the environment, and the economy [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 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 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

 doi:10.20944/preprints202503.2251.v1

due to subjective interpretation [8]. However, with the growing availability of large-scale aviation safety datasets, there is a pressing need for more efficient, automated approaches to accurately infer injury levels from textual incident reports [9]. This task is critical for timely safety assessments. It could benefit significantly from the application of natural language processing (NLP) and deep learning models, which offer scalable and consistent solutions for textual analysis [10,11].

The primary objective of this study is to explore how advanced NLP and deep learning 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 question: To what extent can the injury levels be inferred from the textual narratives using state-of-the-art NLP techniques? Currently, the severity of incidents is often assessed manually, with experts reviewing reports to classify injury levels, a process prone to inefficiency and inconsistencies [12].

The motivation for this study arises from the challenges associated with analyzing the vast amounts of unstructured data generated by the aviation industry. NLP, particularly when combined with deep learning architectures such as Simple Recurrent Neural Network (sRNN), Gated Recurrent Unit (GRU), Bidirectional Long Short-Term Memory (BLSTM), offers a promising solution for automating the extraction of valuable information from these reports [13]. These deep learning 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 based on 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, making traditional analysis methods inadequate [13].

The contributions of this research are twofold. 1) It demonstrates how deep learning 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. By applying advanced NLP and deep learning techniques, this research seeks to enhance the efficiency and accuracy of aviation safety analyses [14].

The remainder of the paper is structured as follows: Section 2 reviews the existing literature on NLP and deep learning 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. Their work underscored the importance of leveraging NLP techniques for improving risk management. However, most studies 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 gap represents a key opportunity for advancing aviation safety research.

Nanyonga et al. [16], applied deep learning 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

 doi:10.20944/preprints202503.2251.v1

subsequent safety actions. Similarly, Zhang et al. [13] demonstrated the effectiveness of LSTM-based models for incident classification, highlighting deep learning's capability 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 recent studies have applied advanced deep learning methods, including RNNs, LSTMs, and GRUs, to aviation safety reports, achieving strong performance in incident classification and trend analysis [17–19]. These models excel in sequential text data analysis, extracting meaningful patterns from complex reports. Despite the promising results, few studies have tackled the specific task of injury level classification, a key element in understanding incident severity and guiding safety measures.

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 [20], but its application to injury level prediction in aviation safety is still in its infancy. This presents an exciting opportunity for further research into its potential in aviation safety analysis.

Ahmad et al. [21] applied deep learning 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 deep learning 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 [22,23].

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 deep learning 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. However, the specific task of classifying injury levels in Australia remains largely unaddressed.

While significant progress has been made in applying deep learning and NLP to aviation safety data, the task of predicting injury levels from textual descriptions remains an underexplored area. Studies have highlighted the promise of deep learning models like RNNs, LSTMs, GRUs, and DistilBERT in improving incident classification and safety analysis. This research aims to fill the gap by investigating how advanced NLP techniques can be applied to predict injury levels from aviation safety reports, thus not only improving the accuracy and scalability of safety analyses but also providing valuable insights for industry stakeholders and regulatory bodies to enhance risk assessment and prevention strategies.

3. Materials and Methods

The study is designed to assess the efficacy of deep learning models in predicting injury levels from unstructured textual narratives in aviation safety reports. To accomplish this, 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, GRU, and DistilBERT. The implementation process, including performance evaluation, followed a systematic procedure to ensure robustness and reliability of the results.

 doi:10.20944/preprints202503.2251.v1

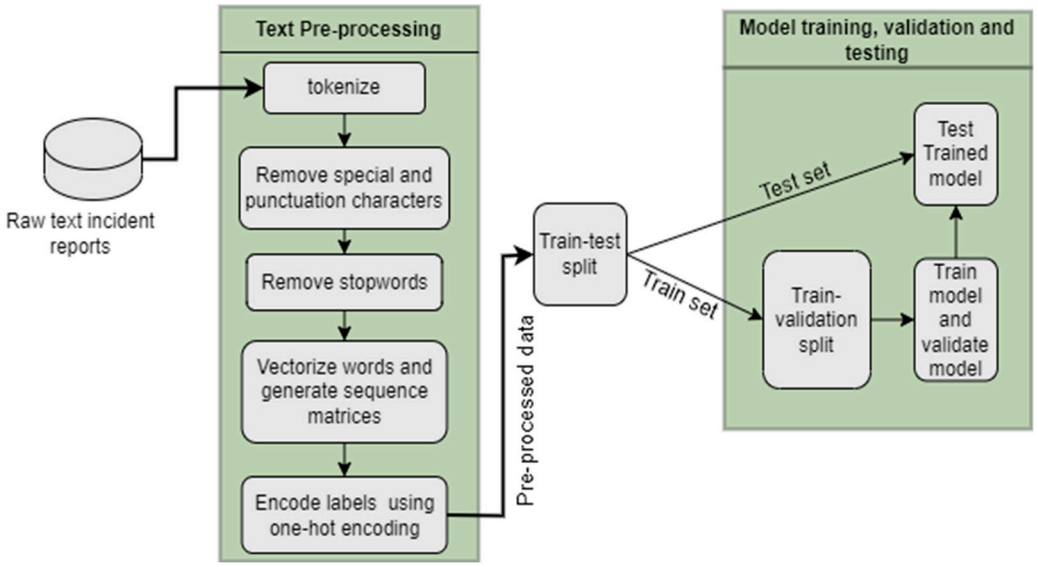

Figure 1. Methodological framework [26].

3.1. Data Collection

The dataset for this study is derived from the Australian Transport Safety Bureau (ATSB) Authorities, consisting of over 50,000 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.

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 library, which provides a comprehensive suite of deep learning 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 utilized 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. Spacy is a powerful Python library tailored for text-processing tasks, including named entity recognition, word tagging, and lemmatization. It 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 enhancing the generalization capability of 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 100,000, allowing for the inclusion of a diverse set of terms and enhancing the

 doi:10.20944/preprints202503.2251.v1

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 [27]. This division ensures that the model is trained on a substantial portion of the data while allowing for unbiased evaluation on the held-out test set.

3.3. Text Classification

To 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 set acted as a reliable benchmark, ensuring that the models were not overfitting to the training data.

In this study, four distinct deep learning architectures were utilized for text classification tasks: GRU, BLSTM, sRNN, and DistilBERT. Each of these models offers unique advantages for processing and classifying text. GRU and BLSTM, which are both variants of RNNs, excel at capturing sequential dependencies and context in textual data, while sRNN provides a simpler, yet effective, architecture for time-series and sequence modeling. DistilBERT, a transformer-based model, offers advanced performance by leveraging pre-trained language representations and self-attention mechanisms, which are particularly well-suited for handling long-range dependencies within text. To address class imbalance in the dataset, class weights were applied across all models during training [28]. 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, selected for its efficiency in gradient-based optimization and its ability to handle sparse gradients and adaptive learning rates. 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 [29]. Future research could explore alternative optimization techniques to further enhance model performance.

3.4. Deep Learning Architecture

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

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 inherent in the textual data. 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. Additionally, ReLU mitigates issues associated with vanishing gradients, promoting efficient training and faster convergence.

The output layer employed the SoftMax activation function, which is particularly well-suited for multi-class classification tasks. 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.

This architecture provided a consistent framework across all models, ensuring reliable and comparable performance metrics. The refined design allowed effective model evaluation, enabling a systematic investigation into the efficacy of each approach. For a visual representation of the deep learning architectures utilized in this study, please refer to Figure 2.

 doi:10.20944/preprints202503.2251.v1

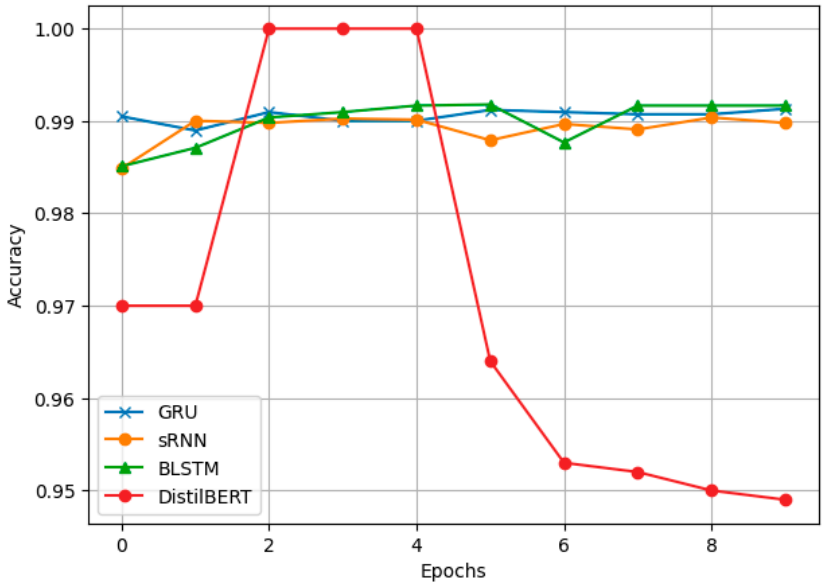

Figure 2. Validation Accuracy of each model.

3.3.1. RNN

The first deep learning architecture considered is the standard RNN, which is 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 [30]. Despite their advantages in handling sequential data, standard RNNs suffer from limitations in capturing long-range dependencies due to the vanishing gradient problem.

3.3.1. BLSTM

To overcome the shortcomings of the standard RNN, the BLSTM network is employed. LSTM, a specialized form of RNN, 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, which allows the model to capture context from both the past and future [31]. This makes BLSTMs particularly effective in understanding the full context of aviation safety narratives. By leveraging the BLSTM model, the study aims to improve the accuracy of injury level classification by considering both previous and future words in the text sequence.

3.3.1. 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 have fewer parameters than LSTMs, making them computationally more efficient while retaining the ability to capture long-range dependencies in the input data [32]. In this study, GRUs are tested alongside LSTMs to determine their effectiveness in classifying injury levels from aviation safety narratives. While GRUs have shown competitive performance in a variety of sequence modeling tasks, their ability to handle the complexity of aviation safety narratives is assessed and compared to other architectures.

3.3.1. DistilBERT

As part of the exploration of cutting-edge NLP models, this study also incorporates DistilBERT, a smaller, faster, and more efficient version of the original BERT model. BERT, or Bidirectional Encoder Representations from Transformers, has set new benchmarks in a wide range of NLP tasks

 doi:10.20944/preprints202503.2251.v1

due to its transformer-based architecture, which captures context from both directions of a given text sequence [33]. DistilBERT is a distilled version of BERT, meaning it retains most of the performance of BERT but with fewer parameters and faster processing time [20]. DistilBERT is tested in this study to assess its ability to classify Injury levels from aviation safety narratives, particularly considering its efficiency and scalability in handling large datasets.

3.5. Model Implementation

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 BLSTM, 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) 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).

3.6. Model Performance Evaluation

In this study, model performance is evaluated in the context of multi-class classification. To ensure a rigorous assessment of the models' effectiveness, a set of well-established performance metrics is employed. These metrics include accuracy, precision, recall, and F1-score, which are standard measures in classification tasks and provide a comprehensive evaluation of the models' ability to classify incidents across multiple injury categories, such as minor, serious, and fatal.

Accuracy serves as a foundational metric, representing 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, which balances precision and recall, is particularly useful in cases of class imbalance, as it prevents a model from favoring the majority class. The F1-score is the harmonic means of precision and recall, offering a more holistic view of model performance.

The performance of the models was evaluated using these metrics, with calculations based on the following formulas as summarized in Table 1: accuracy is the ratio of correct predictions (true positives and true negatives) to the total number of predictions; precision is the ratio of true positives to the sum of true positives and false positives; recall is the ratio of true positives to the sum of true positives and false negatives; and the F1-score is the harmonic mean of precision and recall. 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.

 doi:10.20944/preprints202503.2251.v1

Table 1. Performance Evaluation Metrics.

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

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. Key performance metrics, including accuracy, precision, recall, and F1-score, are employed to assess the reliability and effectiveness of each model. Additionally, validation accuracy and loss trends are examined to evaluate model convergence and generalization.

4.1. Overall Model Performance

Table 2 presents a comparative evaluation of the four models: sRNN, GRU, BLSTM, and DistilBERT, based on key classification performance metrics, including accuracy, precision, recall, and F1-score. 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.

Among the deep learning 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 architectures provide robust classification capabilities, though BLSTM demonstrates slightly better generalization. These findings underscore the effectiveness of deep learning models in handling sequential data, with BLSTM benefiting from its bidirectional processing ability, allowing it to capture contextual dependencies more effectively.

Table 2. 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	**1.00**	**1.00**	**1.00**	**1.00**

4.2. Class-Wise Performance Evaluation

A more detailed evaluation of model performance across individual injury severity categories (Nil, Minor, Fatal, Serious) is provided in Table 3. DistilBERT again demonstrates flawless classification, achieving perfect precision, recall, and F1-score for all injury levels. This result confirms its superior ability to capture nuanced patterns in aviation safety incident narratives.

Among the deep learning 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

 doi:10.20944/preprints202503.2251.v1

recall for Minor injuries, BLSTM achieved the highest score (0.7178), signifying its ability to correctly identify more instances of this category despite its inherent complexity.

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.

Table 3. 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

4.3. Model Validation Performance

To further validate the robustness of the models, Figures 2 and 3 illustrate the validation accuracy and loss trends across training epochs. DistilBERT achieved a perfect accuracy of 100% 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. In contrast, the deep learning models exhibited gradual improvements in accuracy over successive epochs. BLSTM 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 highlights the varying training dynamics and performance characteristics of transformer-based models like DistilBERT in comparison to more traditional deep learning architectures.

As seen in Figure 3, DistilBERT demonstrates a sharp reduction in validation loss early in training, reaching near-zero values. This result indicates its ability to effectively capture linguistic structures with minimal overfitting. Among deep learning models, BLSTM exhibits the lowest validation loss, signifying its superior convergence and improved stability compared to GRU and sRNN. These findings reinforce BLSTM's suitability for handling aviation safety data, as it maintains high accuracy while effectively minimizing classification errors.

 doi:10.20944/preprints202503.2251.v1

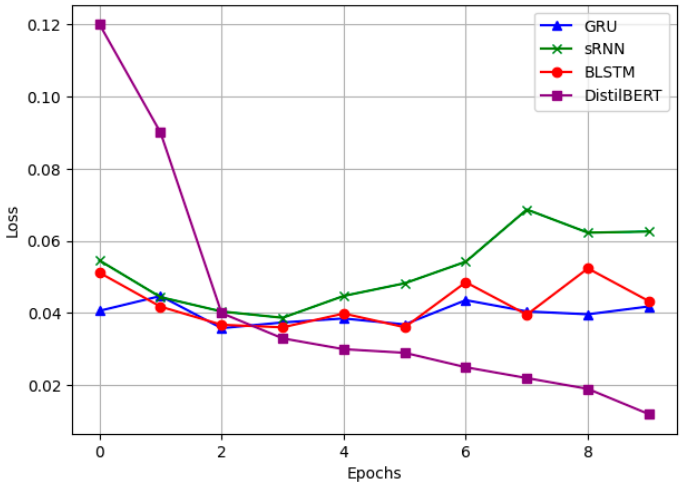

Figure 3. Validation Loss of each model.

4.4. Comparative Analysis of Model Performance

To further assess the classification capabilities of the deep learning models, Figure 4 presents the confusion matrices for BLSTM, the highest-performing deep learning model. This matrix provides an in-depth examination of the model's classification tendencies across injury categories.

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 discrepancy 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, could enhance its classification performance in underrepresented categories.

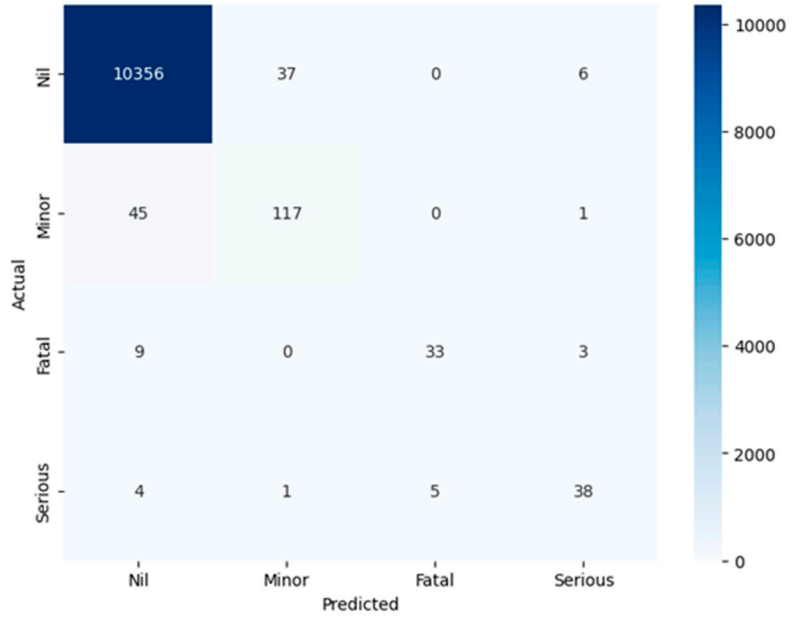

Figure 4. Confusion Matrix for BLSTM.

4. Ablation

To better understand the contributions of various modeling approaches, we conducted an ablation study evaluating the impact of different deep learning architectures and techniques on

 doi:10.20944/preprints202503.2251.v1

classification performance. One key aspect examined was the effect of class weighting, which was 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 [34] or data augmentation methods [35] could further enhance model robustness.

Additionally, the study compared traditional recurrent models sRNN, GRU and BLSTM, against DistilBERT, a transformer-based model. The results demonstrated that while deep learning 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 [36].

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 [37]. 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 natural language processing tasks [38].

4.1. Discussion

The findings of this study demonstrate the efficacy of advanced deep learning models for aviation safety text classification. The results indicate that DistilBERT significantly outperforms traditional recurrent architectures, achieving perfect classification across all injury severity categories. This aligns with prior research highlighting the advantages of transformer-based models in text classification tasks [39,40]. The ability of DistilBERT to maintain high precision, recall, and F1-score suggests that self-attention mechanisms effectively capture long-range dependencies in textual descriptions of aviation incidents, reducing ambiguity in classification.

Among the recurrent models, BLSTM demonstrated the highest accuracy, outperforming sRNN and GRU, 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 [11,41]. However, despite the improved performance, BLSTM and other deep learning 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 deep learning models such as BLSTM struggled to stabilize. This suggests that while deep learning models effectively capture sequential patterns, they may struggle with distinguishing rare but critical cases, warranting further investigation into specialized imbalance-handling strategies.

4.2. 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

 doi:10.20944/preprints202503.2251.v1

or pruning [20]. Further research is needed to explore efficient implementations of these models for large-scale aviation safety monitoring.

Lastly, 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 [42].

5. Conclusion and Future Work

The results of this study demonstrate the effectiveness of advanced deep learning architectures in aviation safety text classification. The comparison of traditional recurrent neural networks and transformer-based models highlights the superior performance of DistilBERT, which achieved perfect classification accuracy across all injury severity categories. The bidirectional structure of BLSTM also exhibited strong performance, outperforming simpler recurrent models such as sRNN and GRU. 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.

Moving forward, future work should focus on enhancing the explainability of AI-driven aviation safety models. The implementation of interpretable machine learning methods, such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations), could improve stakeholder trust and facilitate regulatory acceptance. 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.

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:

BLSTM	Bidirectional Long Short-Term Memory
DistilBERT	Distilled Bidirectional Encoder Representations from Transformers

 doi:10.20944/preprints202503.2251.v1

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 Olja, Gvozdenović Slobodan, Vasov Ljubiša, Mirosavljević Petar %J Technological, economy economic development of. Costs of unsafety in aviation. 2010;16(2):188-201.
2. Somerville Alexander, Lynar Timothy, Wild Graham %J Transportation Engineering. The nature and costs of civil aviation flight training safety occurrences. 2023;12:100182.
3. Harris Don, Li Wen-Chin %J Ergonomics. Using Neural Networks to predict HFACS unsafe acts from the pre-conditions of unsafe acts. 2019;62(2):181-91.
4. Shappell Scott, Detwiler Cristy, Holcomb Kali, Hackworth Carla, Boquet Albert, Wiegmann Douglas A. Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system. Human error in aviation: Routledge; 2017. p. 73-88.
5. 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.
6. Xiong Minglan, Wang Huawei, Wong Yiik Diew, Hou Zhaoguo %J Advanced Engineering Informatics. Enhancing aviation safety and mitigating accidents: A study on aviation safety hazard identification. 2024;62:102732.
7. Slikboer Reneta, Muir Samuel D, Silva S Sandun M, Meyer Denny %J Systematic reviews. A systematic review of statistical models and outcomes of predicting fatal and serious injury crashes from driver crash and offense history data. 2020;9:1-15.
8. 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.
9. Zhang Chenyang, Liu Chenglin, Liu Haiyue, Jiang Chaozhe, Fu Liping, Wen Chao, Cao Weiwei %J Aerospace. Incorporation of pilot factors into risk analysis of civil aviation accidents from 2008 to 2020: A data-driven Bayesian network approach. 2022;10(1):9.
10. Nanyonga Aziida, Wasswa Hassan, Joiner Keith, Turhan Ugur, Wild Graham. A Multi-Head Attention-Based Transformer Model for Predicting Causes in Aviation Incident. 2025.
11. Hochreiter S. J. Neural Computation M. I. T. Press. Long Short-term Memory. 1997.
12. Kazi Naumaan Mohammed Saeed. Using Machine Learning Models to Study Human Error Related Factors in Aviation Accidents and Incidents: Dublin, National College of Ireland; 2020.
13. Zhang Xiaoge, Srinivasan Prabhakar, Mahadevan Sankaran %J Safety science. Sequential deep learning from NTSB reports for aviation safety prognosis. 2021;142:105390.
14. Paul Saptarshi, Purkaystha Bipul Syam, Das Purnendu %J International journal of advanced research in computer science. NLP TOOLS USED IN CIVIL AVIATION: A SURVEY. 2018;9(2).
15. Bloedorn Eric, editor Mining aviation safety data: A hybrid approach. Armed Forces Communications and Electronics Association (AFCEA) First Federal Data Mining Symposium, Washington DC; 2000.
16. 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.

 doi:10.20944/preprints202503.2251.v1

17. 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.
18. Zhou Di, Zhuang Xiao, Zuo Hongfu, Wang Han, Yan Hongsheng %J IEEE Access. Deep learning-based approach for civil aircraft hazard identification and prediction. 2020;8:103665-83.
19. Nanyonga Aziida, Wasswa Hassan, Wild Graham, editors. Comparative Study of Deep Learning Architectures for Textual Damage Level Classification. 2024 11th International Conference on Signal Processing and Integrated Networks (SPIN); 2024: IEEE.
20. Sanh Victor, Debut Lysandre, Chaumond Julien, Wolf Thomas %J arXiv preprint arXiv:.01108. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019.
21. Ahmad Istiak, Alqurashi Fahad, Abozinadah Ehab, Mehmood Rashid %J Sustainability. Deep journalism and DeepJournal V1. 0: a data-driven deep learning approach to discover parameters for transportation. 2022;14(9):5711.
22. Rehman Amjad, Saba Tanzila, Mujahid Muhammad, Alamri Faten S, ElHakim Narmine %J Electronics. Parkinson's disease detection using hybrid LSTM-GRU deep learning model. 2023;12(13):2856.
23. Ali Amir R, Kamal Hossam %J Technologies. Time-to-Fault Prediction Framework for Automated Manufacturing in Humanoid Robotics Using Deep Learning. 2025;13(2):42.
24. Zhong Botao, Pan Xing, Love Peter ED, Sun Jun, Tao Chanjuan %J Advanced Engineering Informatics. Hazard analysis: A deep learning and text mining framework for accident prevention. 2020;46:101152.
25. Nanyonga Aziida, Wasswa Hassan, Wild Graham, editors. Topic Modeling Analysis of Aviation Accident Reports: A Comparative Study between LDA and NMF Models. 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON); 2023: IEEE.
26. 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.
27. 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.
28. Gupta Akhilesh, Tatbul Nesime, Marcus Ryan, Zhou Shengtian, Lee Insup, Gottschlich Justin. Class-weighted evaluation metrics for imbalanced data classification. 2020.
29. Kingma Diederik P %J arXiv preprint arXiv:. Adam: A method for stochastic optimization. 2014.
30. Salem Fathi M %J arXiv preprint arXiv:.09022. A basic recurrent neural network model. 2016.
31. Schuster Mike, Paliwal Kuldip K %J IEEE transactions on Signal Processing. Bidirectional recurrent neural networks. 1997;45(11):2673-81.
32. Dey Rahul, Salem Fathi M, editors. Gate-variants of gated recurrent unit (GRU) neural networks. 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS); 2017: IEEE.
33. 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.
34. Ross T-YLPG, Dollár GKHP, editors. Focal loss for dense object detection. proceedings of the IEEE conference on computer vision and pattern recognition; 2017.
35. Zhu Xianglei, Men Jianfeng, Yang Liu, Li Keqiu %J International Journal of Machine Learning, Cybernetics. Imbalanced driving scene recognition with class focal loss and data augmentation. 2022;13(10):2957-75.
36. Vaswani A. J. Advances in Neural Information Processing Systems. Attention is all you need. 2017.
37. Khandelwal Urvashi, Clark Kevin, Jurafsky Dan, Kaiser Lukasz %J arXiv preprint arXiv:.08836. Sample efficient text summarization using a single pre-trained transformer. 2019.
38. Devlin Jacob %J arXiv preprint arXiv:.04805. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.
39. Brown Tom, Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared D, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda %J Advances in neural information processing systems. Language models are few-shot learners. 2020;33:1877-901.

 doi:10.20944/preprints202503.2251.v1

40. Lan Zhenzhong, Chen Mingda, Goodman Sebastian, Gimpel Kevin, Sharma Piyush, Soricut Radu %J arXiv preprint arXiv:.11942. Albert: A lite bert for self-supervised learning of language representations. 2019.
41. Graves Alex, Mohamed Abdel-rahman, Hinton Geoffrey, editors. Speech recognition with deep recurrent neural networks. 2013 IEEE international conference on acoustics, speech and signal processing; 2013: Ieee.
42. 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.