Student: Anthony Eddei Kwofie

Programme: MSc Artificial Intelligence and Data Science

Module: Applied Artificial Intelligence

Submission Date: 18th December 2025

Title:

Automatic Hospitalisation Risk Classification in OSHA Incident Narratives Using Traditional NLP Techniques and Deep Learning Approaches

1. Problem Definition and Scope

Workplace accident reports typically contain short narrative descriptions of the incident and its outcome. These narratives include consistent severity cues. Terms such as fractured, burns, or transported are strongly associated with serious injuries, while phrases like minor cut or returned to work usually indicate lower severity. Because these patterns recur across reports, they are well suited to learning using Natural Language Processing (NLP).

OSHA publishes thousands of such reports each year, making manual review time-consuming. Automating an initial screening step could help safety teams prioritise potentially serious cases earlier.

This project developed a binary text classifier to predict hospitalisation using the narrative alone. Two traditional machine-learning models and two deep-learning models were evaluated using accuracy, recall, precision, and F1 score, with class imbalance addressed where appropriate. Logistic Regression with SMOTE achieved the strongest overall F1 score and offered the best balance of performance, interpretability, and computational efficiency for deployment.

2. Importance and Background Review

Text classification is widely used in NLP applications such as spam detection, sentiment analysis and clinical text processing. The same techniques are highly relevant to workplace safety, where large volumes of accident reports are still reviewed manually. Although OSHA publishes extensive injury data each year (OSHA, 2025), there is no widely deployed system that predicts hospitalisation risk directly from free-text narratives. Automating this task could support earlier identification of serious incidents and more efficient allocation of safety resources.

Previous NLP studies show that, short accident and medical narratives contain strong linguistic patterns that can be learned using traditional machine-learning models such as Naïve Bayes and Logistic Regression. These models perform well with sparse representations like TF-IDF and remain popular because they are fast, interpretable and easy to deploy using tools such as scikit-learn (Pedregosa et al., 2011). However, they rely largely on word frequency and often fail to capture contextual meaning.

More recent research has demonstrated the effectiveness of deep-learning models for text understanding. LSTM networks capture sequential dependencies in language (Hochreiter and Schmidhuber, 1997), while transformer-based models such as BERT provide contextual embeddings that improve performance across many NLP tasks (Devlin et al., 2019). Lightweight variants such as DistilBERT offer similar representational power at lower computational cost (Reimers and Gurevych, 2019).

Despite these advances, most occupational-safety NLP research focuses on report categorisation or hazard detection rather than direct hospitalisation prediction. In addition, traditional and deep-

learning models are often evaluated separately. This project addresses that gap by systematically comparing both approaches on the same dataset.

The relevance of this problem is reinforced by global safety statistics. The International Labour Organization estimates that, approximately 2.9 million work-related deaths occur annually worldwide, highlighting the need for scalable tools that can rapidly identify high-risk cases in large accident databases.

3. SMART Objectives

To guide the project, the following SMART objectives were defined:

1. **Specific:** Develop a binary classifier that predicts whether an accident resulted in hospitalisation using only the narrative text.
2. **Measurable and Achievable:** Achieve an F1 score of at least 0.90, exceeding the majority-class baseline (0.89) and consistent with results reported in related NLP studies using TF-IDF and imbalance handling.
3. **Relevant:** Support workplace safety decision-making by enabling early identification of potentially severe incidents.
4. **Time-bound:** Complete all experiments and reporting by 18 December 2025.

4. Dataset Description and Suitability

The dataset comprised 79,351 OSHA injury reports (2015–2022) sourced from the Kaggle OSHA Severe Injury Reports collection. Although each record included structured fields, this study used only the Final Narrative as input text. The Hospitalized field was recoded into a binary target variable.

No reports were missing narrative text, leaving all records available for analysis. The dataset was class-imbalanced, with hospitalised cases forming the majority, but narratives spanned multiple industries and injury types, providing sufficient diversity for model learning.

4.1 Dataset Preprocessing

Narratives were lowercased, stripped of punctuation and line breaks, and normalised by collapsing extra whitespace. Empty texts were removed. Data was split 80/20 using stratified sampling (random_state = 42).

To address imbalance, SMOTE was applied only to the training set. Traditional models used TF-IDF features, while deep-learning models used tokenised and padded sequences for fixed-length input.

5. Exploratory Data Analysis and Baseline

I explored the narratives to understand their structure and assess whether they contained learnable severity signals.

5.1 Narrative Length

Most reports were short, typically 10–60 words, with an average of about 33 words. A small subset exceeded 150 words, usually describing complex incidents.
Based on this distribution, I set a maximum sequence length of 150 tokens for the BiLSTM and DistilBERT models to capture most narratives without unnecessary computation.

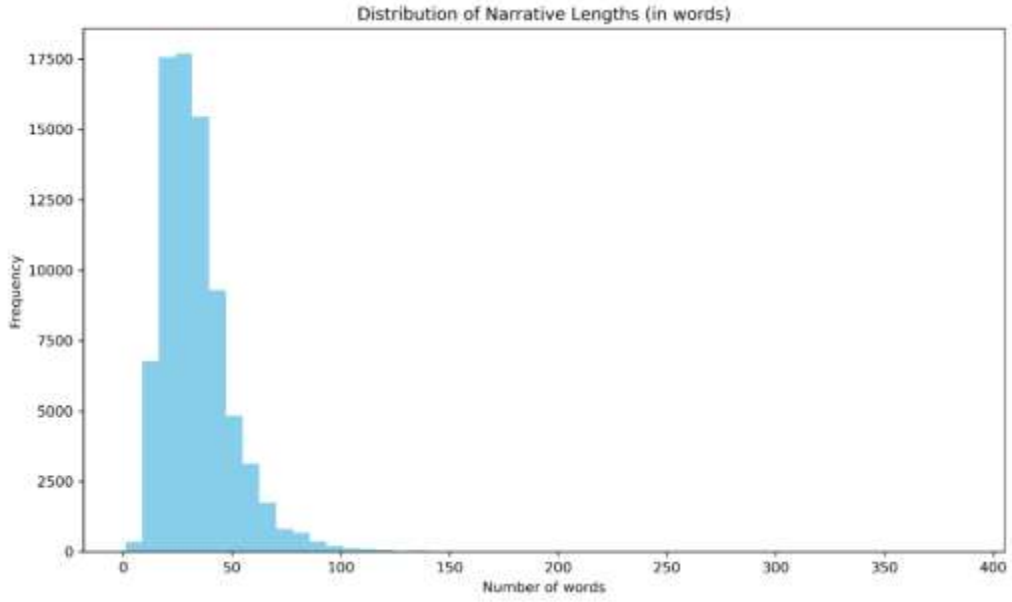

Figure 1: *Distribution_Narrative_Lengths*

5.2 Common Vocabulary

I examined the most frequent words across all narratives. Terms such as employee**,** injury**,** hand**,** machine, and fell dominated, clearly reflecting industrial environments, machinery use, and manual work activities.

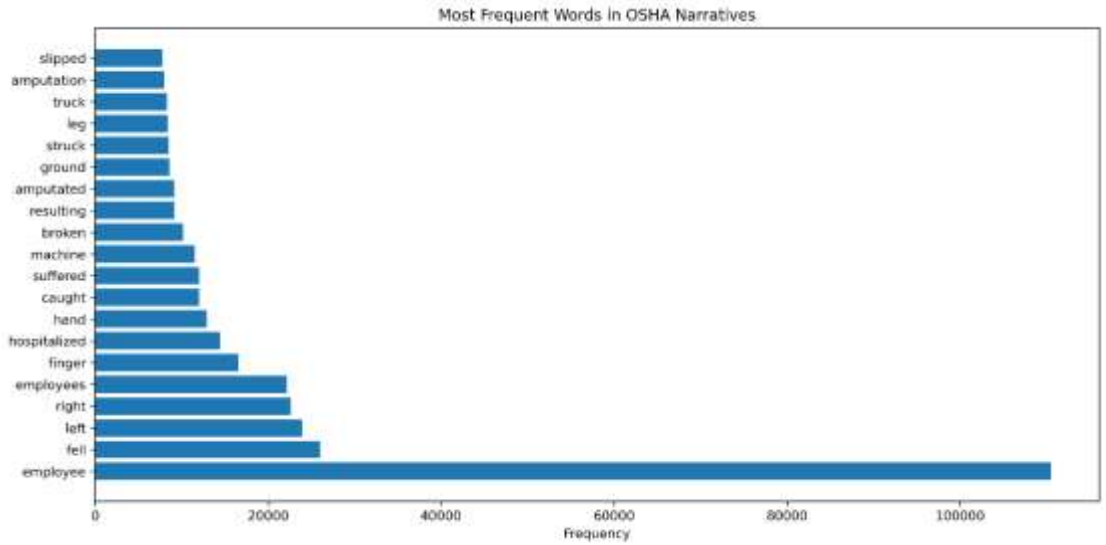

Figure 2: *Top_Keywords*

5.3 Word Clouds

I generated separate word clouds for hospitalised and non-hospitalised cases. Hospitalised narratives emphasised terms such as fractured**,** crushed**,** burns, and transported, while non-hospitalised cases focused on minor injuries and first-aid treatment.

.

Figure 3: *Wordcloud_Hospitalized*

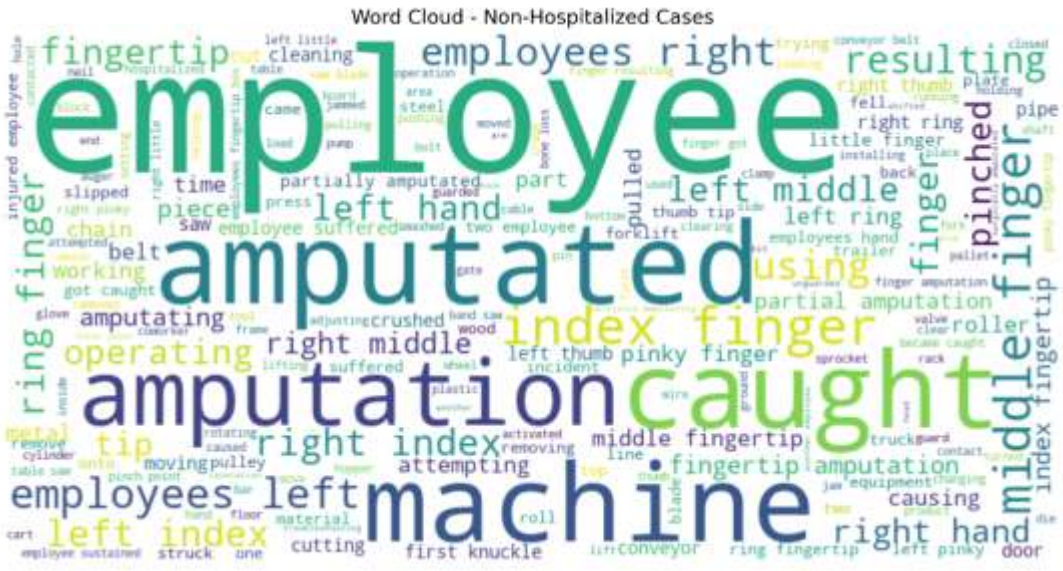

Figure 4: *Wordcloud_Non_Hospitalized*

These observations suggested that the narratives contained strong linguistic indicators of severity, meaning NLP classification was likely to succeed.

6. Baseline Model

I used a majority-class baseline that predicted hospitalisation for every case. It achieved 0.803 accuracy and an F1 score of 0.891 due to perfect recall on the majority class. However, it identified no non-hospitalised cases, producing zero true negatives. As a result, the baseline was operationally useless but provided a clear benchmark that all trained models needed to exceed.

7. Traditional Machine Learning Methods

7.1 Naïve Bayes

Multinomial Naïve Bayes trained on TF-IDF features achieved 0.882 accuracy and an F1 score of 0.925, confirming that simple probabilistic models can capture severity-related word patterns. However, applying SMOTE reduced performance (F1 = 0.906), indicating that synthetic samples disrupted Naïve Bayes' independence assumptions in sparse text space.

7.2 Logistic Regression

Logistic Regression outperformed Naïve Bayes, achieving 0.906 accuracy and an F1 score of 0.941. When combined with SMOTE, performance improved further (accuracy 0.911, F1 0.942, precision 0.981). Unlike Naïve Bayes, Logistic Regression handled oversampling well, maintaining stable decision boundaries and strong generalisation, making it the best overall traditional model and suitable for deployment.

.

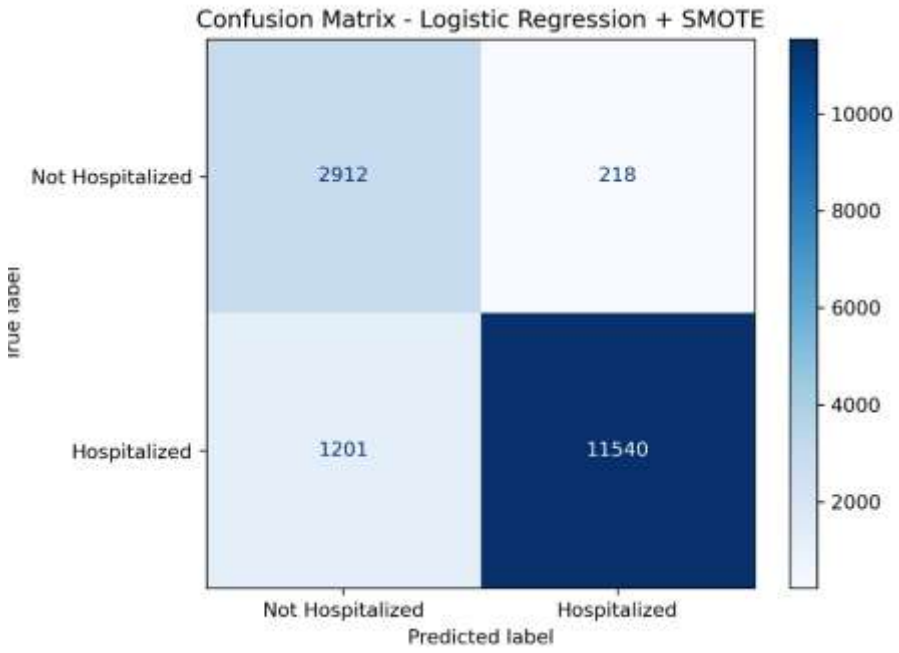

Figure 5: *Confusion_Matrix_Log_Smote*

Model performance was summarised using confusion matrices, ROC curves and precision–recall curves, which demonstrated strong class separation, as evidenced by consistently high ROC-AUC values and well-shaped precision–recall curves:

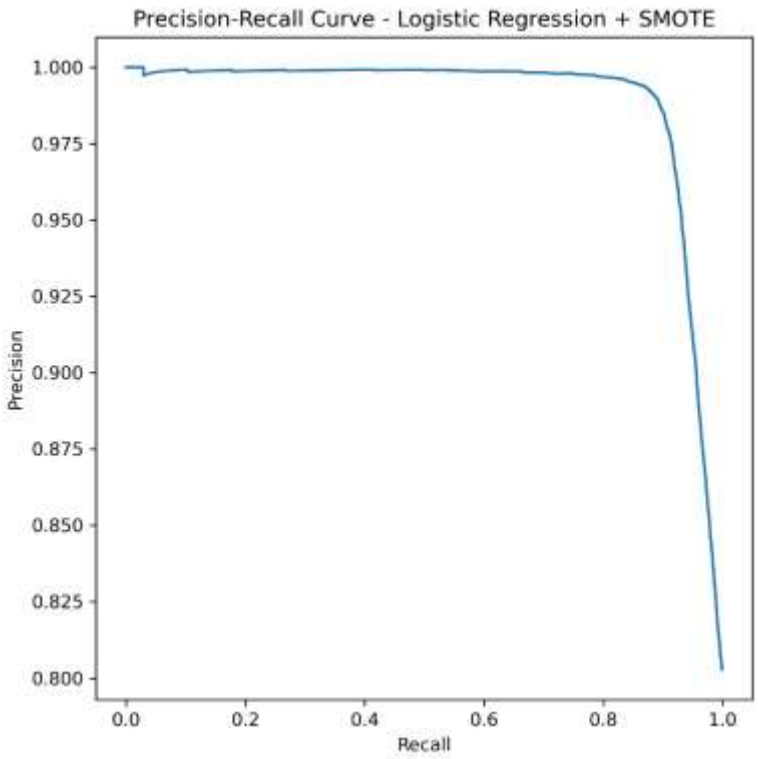

Figure 6: *Roc_Log_Smote*

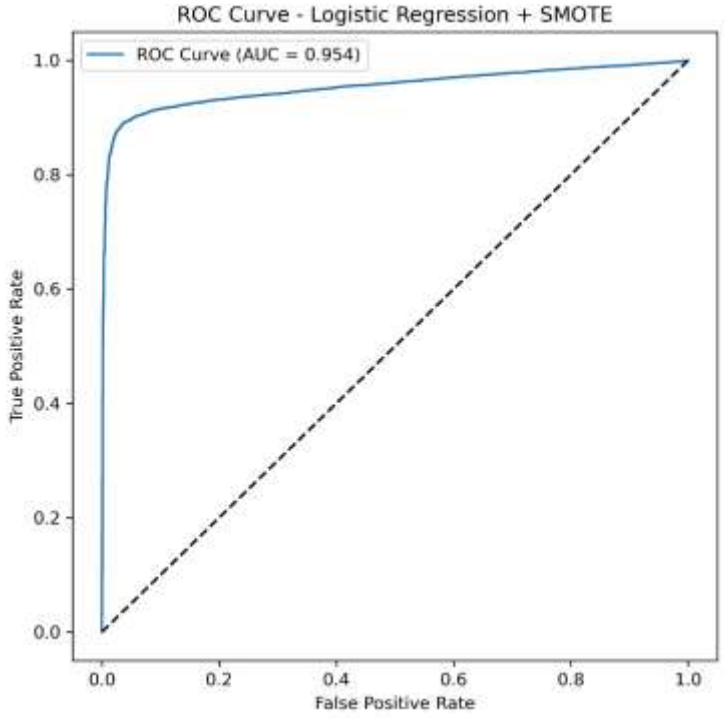

Figure 7: *Pr_Log_Smote*

7.3 Traditional Model Exclusions and Selection

Support Vector Machines (SVM) were reviewed because of their strong performance in text classification. However, training an SVM on a large TF-IDF matrix with thousands of features is computationally expensive and requires extensive tuning, making it impractical for this project.

Random Forest was also considered, but tree-based models perform poorly on very sparse TF-IDF features and scale inefficiently as dimensionality increases. Given the size and sparsity of the dataset, this approach was not suitable.

k-Nearest Neighbours (kNN) was excluded due to poor scalability and sensitivity to noise in short narratives. With over 79,000 samples, inference would be slow and unstable.

Overall, **Naïve Bayes and Logistic Regression** were selected because they trained quickly, handled sparse text features effectively, and produced interpretable outputs—an important requirement in safety-critical applications.

8. Deep Learning Methods

8.1 Bidirectional LSTM

The Bidirectional LSTM achieved 0.899 accuracy and an F1 score of 0.936 without imbalance handling. Introducing class weighting improved performance (accuracy 0.907, F1 0.940, precision 0.986). Training curves showed stable convergence, and limiting training to three epochs reduced overfitting. Class weighting improved recognition of non-hospitalised cases while maintaining high precision. Although effective, the model trained more slowly and ranked slightly below Logistic Regression with SMOTE.

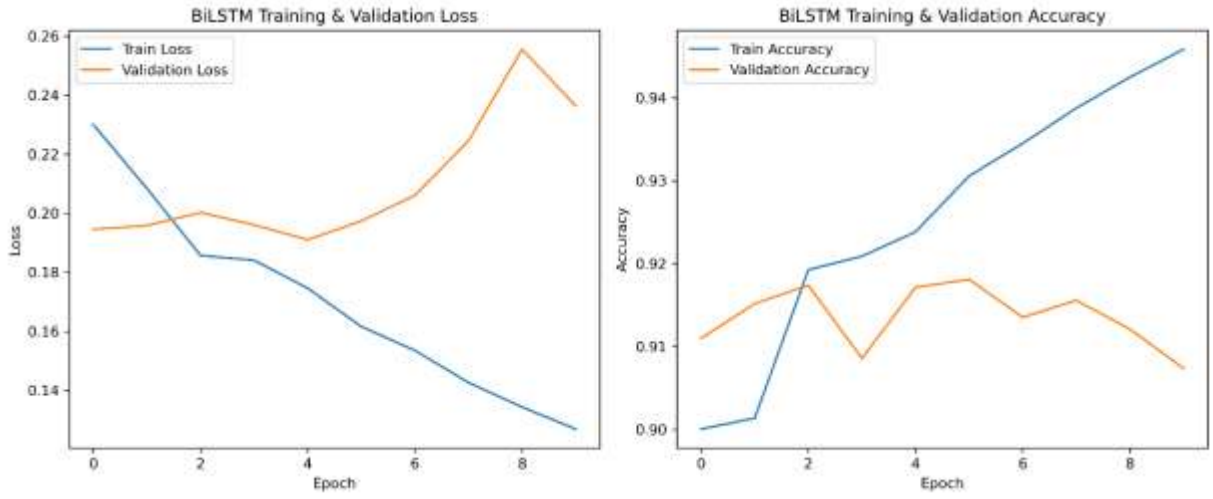

Figure 8: *Bilstm_Training_Curves*

Although strong, this model still ranked slightly below Logistic Regression + SMOTE.

8.2 DistilBERT

DistilBERT sentence embeddings combined with Logistic Regression achieved **0.903 accuracy** and an F1 score of 0.939 without transformer fine-tuning. Lowering the decision threshold from 0.50 to 0.45 improved recall for non-hospitalised cases (0.938) with a small drop in accuracy. This showed that threshold adjustment can improve fairness in imbalanced settings without modifying transformer weights. Full fine-tuning was not performed due to hardware limits.

8.3 Deep Learning Models Not Implemented

A basic RNN was reviewed but excluded because it struggles with long-range dependencies and can lose early information in longer narratives.
A GRU was considered but not implemented, as its expected performance was similar to the BiLSTM without offering clear advantages.
A Text-CNN was also reviewed but excluded because it captures short patterns well but fails to model longer narrative context common in OSHA reports.

8.4 Deep Learning Model Selection

The BiLSTM and DistilBERT were selected as complementary deep learning approaches. The BiLSTM captured bidirectional sequence information, while DistilBERT provided strong semantic representations with manageable computational cost. Other models were either less suitable or redundant.

9. Implementation, Refinement and Evaluation

SMOTE was applied only to TF-IDF-based models, as it is incompatible with sequence networks and transformer embeddings. Deep learning models instead used class weighting or probability threshold adjustment to address class imbalance.

Model	Accuracy	Precision	Recall	F1
Baseline (Majority Class)	0.803	0.803	1.000	0.891
Naive Bayes	0.882	0.943	0.908	0.925
Naive Bayes + SMOTE	0.860	0.978	0.844	0.906
Logistic Regression	0.906	0.949	0.934	0.941
Logistic Regression + SMOTE	0.911	0.981	0.906	0.942
BiLSTM	0.899	0.954	0.919	0.936
BiLSTM + CW	0.907	0.986	0.897	0.940
DistilBERT	0.903	0.947	0.932	0.939
DistilBERT + Thr	0.900	0.937	0.938	0.938

Model performance was assessed using confusion matrices and ROC and precision–recall curves, meeting the requirement for multiple evaluation visualisations.

9.1 Hyperparameter Tuning Summary

I applied light, targeted tuning to improve stability and generalisation: regularisation for classical models, training configuration for the BiLSTM, and probability-threshold adjustment for DistilBERT. Naïve Bayes and Logistic Regression were tuned in scikit-learn, the BiLSTM was implemented in Keras, and DistilBERT was used as a fixed embedding model without transformer fine-tuning.

Model	Tuned Parameter	Tested Values	Selected Value	Effect
Logistic Regression	C	0.1, 1.0, 3.0, 5.0	3.0	Reduced underfitting without overfitting
Naïve Bayes	Alpha	0.5, 1.0, 2.0	1.0	Higher values reduced minority sensitivity
BiLSTM	Epochs	3, 10	3	Shorter training reduced overfitting with class weighting
BiLSTM	Batch size	32, 64, 128	64	Improved convergence stability
DistilBERT + LR	Decision threshold	0.50, 0.45, 0.40	0.45	Improved minority recall

Small, targeted adjustments were sufficient to achieve stable and competitive performance across all models.

10. Discussion

All models outperformed the baseline, confirming that OSHA narratives contain sufficient signal for hospitalisation prediction. Logistic Regression with SMOTE achieved the highest F1 score, likely because short narratives are well suited to TF-IDF features. Naïve Bayes degraded under SMOTE, while Logistic Regression remained stable.

The BiLSTM and DistilBERT models improved with class weighting and threshold adjustment respectively but required more computation. Most errors occurred in very short or ambiguous narratives.

11. Limitations

This study focused only on hospitalisation as a severity proxy; modelling specific injury types would require multi-class or multi-label labels. Narrative quality varied, and short or vague reports reduced reliability. DistilBERT was used only as an embedding model due to hardware limits. Results reflect US reporting styles and may not generalise without retraining.

12. Ethical and Data Governance Considerations

The dataset is public and anonymised, but narrative text still requires careful use. Misclassifying severe cases poses risk, so recall for serious incidents was prioritised. Logistic Regression offers transparency, while neural models would require explainability tools. Any deployment should include human oversight and ongoing monitoring.

13. Conclusion

This project demonstrated that accident narratives can support accurate hospitalisation risk prediction. Logistic Regression with SMOTE achieved the best balance of performance, interpretability, and efficiency. Deep learning models also performed well but required greater computational cost. Deployment should prioritise recall, explainability, and human review.

14. Deployment Architecture

A lightweight deployment pipeline was designed using Logistic Regression due to its strong performance and low computational cost. The model can be served via a REST API and integrated into existing reporting systems, providing real-time risk scores to support safety decision-making.

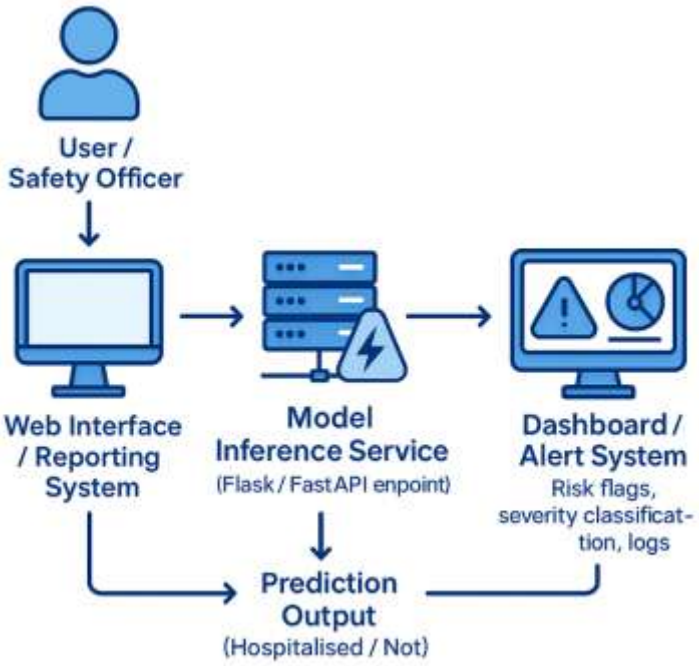

Figure 10: *Proposed deployment architecture for hospitalisation risk classification from OSHA narratives.*.

Deployment Workflow (Operational Use)

In practice, a safety officer submits an incident narrative through a reporting form. The text is sent to an inference service (e.g. Flask or FastAPI), where the trained Logistic Regression model and TF-IDF vectoriser are loaded. The service returns a risk probability that feeds into a dashboard to flag high-risk cases for review.

Logistic Regression provides near-instant predictions on standard hardware. Deep learning models are feasible but better suited to cases requiring deeper semantic analysis. Human oversight remains essential, supported by explainability tools and periodic retraining to manage performance drift.

15. Referencing Style Statement

This report follows the Hull Harvard referencing style, the standard at the University of Hull. Its author–date format supports clear, readable scientific writing and allows sources to be traced without disrupting the narrative, making it appropriate for this technical report.

16. Use of Generative AI

A generative AI tool, specifically ChatGPT, was used solely for research support. All research design, modelling, experimentation, evaluation, and interpretation were carried out independently by the author.

17. References

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *Proceedings of NAACL-HLT*, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Computation*, 9(8), pp. 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

International Labour Organization (ILO) (2023) *Global estimates of occupational accidents and work-related illnesses 2019–2023*. Geneva: ILO. Available at: https://www.ilo.org (Accessed: 1 December 2025).

Kingma, D.P. and Ba, J. (2015) 'Adam: A method for stochastic optimization', *Proceedings of the International Conference on Learning Representations (ICLR)*. Available at: https://arxiv.org/abs/1412.6980 (Accessed: 1 December 2025).

Kaggle (2023) *OSHA Severe Injury Reports (2015–2022)*. Available at: https://www.kaggle.com/datasets/krist0phersmith/osha-severe-incident-reports (Accessed: 1 December 2025).

Occupational Safety and Health Administration (OSHA) (2025) *Occupational Safety and Health Administration (United States) – Federal Injury Data*. Available at: https://www.osha.gov (Accessed: 1 December 2025).

Pedregosa, F. et al. (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Reimers, N. and Gurevych, I. (2019) 'Sentence-BERT: Sentence embeddings using Siamese BERT-networks', *Proceedings of EMNLP*, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

Vaswani, A. et al. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems (NeurIPS)*. doi: 10.48550/arXiv.1706.03762.