SafeAeroBERT: Towards a Safety-Informed Aerospace-Specific Language Model

Sequoia R. Andrade[1]
HX5 LLC, NASA Ames Research Center, Moffett Field, CA, 94053, United States

Hannah S. Walsh[2]
NASA Ames Research Center, Moffett Field, CA, 94053, United States

As aviation systems continue to operate with high traffic, large amounts of documents containing safety-relevant data continue to be generated via reporting systems such as the Aviation Safety Reporting System (ASRS). Advanced natural language processing techniques, specifically pre-trained language models, have shown great success in domain-specific applications; however, the text in aviation safety reports is inundated with jargon and thus not fully utilized by general pre-trained models. In this research, we work towards developing a safety-informed aerospace-specific language model by pre-training a Bidirectional Encoder Representations from Transformer (BERT) model on reports from the Aviation Safety Reporting System and the National Transportation Safety Board. The resulting model, called SafeAeroBERT, is fine-tuned for the specific task of document classification, and can be further tuned for named-entity recognition, relation detection, information retrieval, and summarization. Results from the classification task are compared between SafeAeroBERT, the base BERT, and SciBERT models and show SafeAeroBERT outperforms the general BERT and SciBERT on classifying reports about weather and procedure. SafeAeroBERT can be used on custom tasks, not limited to document classification, and is intended to aid an intelligent knowledge manager for safety report repositories.

I. Nomenclature

ASRS	=	Aviation Safety Reporting System
BERT	=	Bidirectional Encoder Representations from Transformer
IR	=	Information Retrieval
NER	=	Named Entity Recognition
NLP	=	Natural Language Processing
NTSB	=	National Transportation Safety Board
QA	=	Question Answering
RE	=	Relation Extraction

II. Introduction

Aviation regulatory requirements generate large corpora of documents for many different use cases, such as compliance documents, mishap safety reports, and requirements for new system designs. These data sources harbor a vast amount information, yet conventionally this information can only be extracted by trained analysts. The jargon in aviation documents limits the effectiveness of state-of-the-art machine learning models, which may not fully represent the meaning and context of aviation specific concepts. Thus, there is a need for an aviation-specific language model which can effectively ingest documents with aviation jargon and perform natural language processing tasks.

A growing body of research has demonstrated the effectiveness of pre-training Bidirectional Encoder Representations from Transformers (BERT) models [1] on domain-specific text heavy with jargon. BERT models

[1] Research Engineer, sequoia.r.andrade@nasa.gov, funded Under Prime Contract No. 80ARC020D0010 with the NASA Ames Research Center.
[2] Computer Engineer, Intelligent Systems Division, Hannah.s.walsh@nasa.gov

have not only been successfully trained for various domain-specific use cases (e.g., biomedical [2], legal, finance), but also show superior performance on nearly all natural language processing (NLP) tasks. Additionally, transfer learning from using pre-trained models decreases the computational power and training time for downstream tasks [1].

In this research, we define the requirements for an aviation safety BERT model. We pretrain a BERT model on a large corpus of aviation safety documents, fine-tune the model for traditional NLP tasks applied to aviation, and evaluate the performance of the custom model by comparing it to the base BERT models on downstream tasks. As a result, we produce a safety-informed aviation specific language model, named SafeAeroBERT. The resulting domain-specific BERT model pre-trained on aviation safety documents can be used as a backbone for NLP based tools exploring aviation documents or for individual tasks requiring a language model.

III. Background

In this section, we provide an overview of natural language processing, discuss common tasks performed with NLP, and explain the state-of-the-art accomplished by pre-trained language models. Next, existing domain-specific pre-trained language models are identified and their performance on NLP tasks is considered. Finally, we discuss NLP in aerospace applications.

A. Natural Language Processing

Natural language processing is a category of machine learning algorithms centered around natural language understanding and natural language generation [3]. For natural language generation, a typical use-case involves a user interacting with a computer application where the application is known as the speaker [3]. A user inputs information, and the application must respond using natural language [3], such as when an iPhone user asks Siri a question and Siri responds. Natural language understanding encompasses a larger swath of tasks, including the broad categories of phonology (sounds of words), morphology (parts of words), syntax (parts of speech), semantics (meaning of text), and pragmatics (implied meaning of text) [3, 4]. For processing large collections of text reports with the goal of information extraction, semantic methods are primarily used. While there is a large set of sematic methods, those of interest in this research are: document classification [4], named-entity recognition [5], relation extraction [6], question answering [4], information retrieval [4], summarization [4]. A brief description of each is provided below:

1. Classification: Classification is one of the most used NLP tasks across a variety of domains. Classification has been applied to ASRS documents widely in the past with the goal of identifying a documents category based on the text input.
2. Named-entity recognition (NER): Named-entity recognition is a sub-method of information extraction which classifies individual words or sequences of words as entity types [3, 5]. For example, “California”, “United States”, and “Los Angeles” are all locations. The base BERT model is trained on location, person, organizations, and miscellaneous [1].
3. RE: Relation extraction is also a sub-method of information extraction, which detects the relationship between entities and text chunks [3, 6]. A specific subset of RE that is relevant to accident analysis is causality mining (CM) [7]. CM identifies causal relationships between entities and has previously been used in biomedical applications [6, 2].
4. Information retrieval (IR): Information retrieval provides users with relevant documents based on a key word query search or similar document search.
5. Question answering (QA): Though seemingly similar to information retrieval, question answering provides a definite answer to a user question rather than a document. QA is typically applied within a single document, rather than to the entire set of documents like in IR. For example, a user could ask “what was the aircraft type involved in the accident?” in a selected report.
6. Summarization: Summarization can be performed in two methods: abstractive and extractive. Abstractive summarization analyzes the entire document and provides a summary of specified length based on the main ideas, which may or may not contain verbatim sentences from the original text. In contrast, extractive text compiles a summary only from exact sentences within the original text. This task is especially useful for corpora with long documents that may take a user an extended time to read.

Over the course of its existence, various methods have been used for NLP, including rule-based models, lexicons, neural networks, word embeddings, and most recently, pre-trained language models [3]. Pre-trained language models achieve superior performance when compared to conventional bag-of-words and word embedding models across NLP tasks due to the high volume of training data [4]. These models are trained on millions of general documents, which allow them to learn language via the context of words [5]. Consequently, pre-trained models can be applied to downstream tasks, such as document classification and named-entity recognition, through transfer learning [8]. In recent years, pre-trained models have exploded in popularity, resulting in a collection of models with different deep learning architectures trained on various datasets [8]. Two widely used models are Bidirectional Encoder Representations from Transformers (BERT) [1] and Generative Pre-trained Transformer (GPT) [9]. While GPT has numerous benefits, including having a larger corpus used for pre-training and better performance with less training data, GPT is commercially available, rather than open sourced, and only learns text from left to right rather than bi-directionally [9, 10]. Therefore, BERT is chosen for developing the proposed aviation safety language model due to, ease of training, open-source access, and bi-directional left-to-right and right-to-left language model [1]. There are also numerous variants of BERT, such as a siamese implementation called Sentence-BERT (SBERT) [11], a light-weight version known as distilBERT [12], and a further pre-trained version named RoBERTa [13]. Additionally, BERT models have a precedent of success for domain-specific language models via additional pretraining.

B. Domain-Specific Language Models

Due to the open-source nature of BERT models and their distinguished performance across NLP tasks, in recent years there has been an explosion of highly successful domain-specific BERT models. Generally, a domain-specific BERT model is created by further pre-training an existing BERT model on a large collection of domain-specific documents. Some domain-specific BERT models are described in Table **1** and include BioBERT [2], LegalBERT [14], FinBERT [15], ArcheoBERTje [16], SciBERT [17], AeroBERT [18], SpaceTransformers [19], and our proposed model, SafeAeroBERT.

Table 1: Domain-specific BERT models, training information, and accessibility.

Name	Domain	Training data source	Training Data Size	Base model	Accessibility
bioBERT	Biomedical	Pubmed abstracts + full text articles	87 GB	BERT	Open Source
LegalBERT	Law	EU legislation UK legislation European Court of Justice Cases European Court of Human Rights Cases US Court cases US contracts	12 GB	BERT	Open Source
FinBERT	Finance	Routers TRC2 Financial Financial Phrase Book	61 GB	BERT	Open Source
ArcheoBERTje	Archeology	DANS archive excavation reports	2 GB	BERTje	Open Source
SciBERT	Science	Semantic Scholar full text articles	15.4 GB	BERT	Open Source
AeroBERT	Aerospace	NASA NTRS documents	4.3 million records	Unknown	Proprietary
SpaceTransformers	Astronautics	Journal Articles Books Wikipedia	14.3 GB	BERT RoBERTa SciBERT	Open Source
SafeAeroBERT	Aviation Safety	ASRS reports, NTSB reports	1 GB	BERT	Open Source

BioBERT, one of the first domain-specific BERT models released in 2019, is trained on a corpus of biomedical publications and was found to perform better than the generalized BERT model on NER, RD, and QA [2]. LegalBERT includes three models, one using the existing BERT vocabulary, another with a new vocabulary learned from the training corpus, and a third lightweight model [14]. The family of models was evaluated on text classification and NER, with results showing improved performance over the general BERT [14]. FinBERT was pre-trained on financial text, and also outperforms the general BERT model on text classification and sentiment analysis when fine-tuned [15]. ArcheoBERTje is a Dutch archeology-specific BERT model pre-trained and fine-tuned for NER, which again demonstrates that the domain-specific pre-training makes a considerable difference in performance on NER [16]. Envisioned as a language model for the general science domain, SciBERT is pre-trained on scientific journal articles and evaluated on downstream tasks after fine-tuning, NER, text classification, and relation detection (called relation classification by the authors) [17]. SciBERT outperforms the base BERT model on scientific tasks, and the authors find that using a domain-specific vocabulary, rather than the base vocabulary, impacts performance [17]. SpaceTransformers consists of three different models (space-BERT, space-roBERTa, spaceSciBERT) that were further pre-trained on 14.3 GB of space-relevant text found in books, academic abstracts, and Wikipedia pages [19]. The resulting models were then fine-tuned on a dataset of ESA requirements for performing NER (note the researchers call it concept recognition), with the SpaceRoBERTa model performing best on the majority of entities [19].

Most relevant to this research, with the aid of NVIDIA, Rolls Royce has developed AeroBERT [18], an aerospace specific BERT model, by further pre-training an existing model on the complete corpus from the NASA Technical Reports Server (NTRS). However, AeroBERT is proprietary and not accessible to the public. Additionally, the training data from NTRS is not exclusively aviation documents, as NASA also has missions and publications for other domains, such as space exploration, astrobiology, and earth sciences. While AeroBERT may provide an aerospace language model, it is not inherently safety-centric nor open source. Instead, our proposed SafeAeroBERT model is trained exclusively on aviation safety reports and is open source for reuse on aviation NLP tasks.

C. **Natural Language Processing in Aviation**

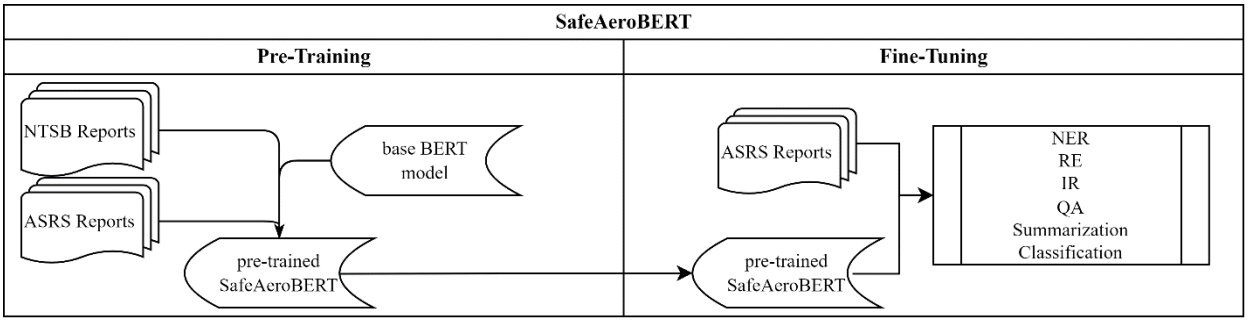
Natural Language Processing is increasingly being used to process text-based documents in aviation [20]. There is a large body of research for classification of ASRS reports using various methods, including using the universal language model for embeddings prior to applying an LSTM classifier for primary and secondary cause identification [21], using a hybrid SVM and neural network model for report severity classification [22], and using clustering to identify driving factors of reports [23] . Other research has used topic modeling to examine trends in ASRS reports [24, 25, 26], while we have previous used topic modeling to identify hazards in aerial wildfire response mishap reports [27]. Named-entity recognition is increasingly being applied to aviation. Researchers have used NER to assist in the digitalization of NOTAMs [28], as well as fine-tuned a base BERT model to detect the high-level aviation entities of system, value, date time, organization, and resource [29]. We have previously used NER to semi-automatically extract failure modes and effects analysis components from wildfire UAS mishap reports [30]. DistilBERT has been applied to ASRS reports for question answer tasks, with around a 70% accuracy rate [31]. Summarization has been applied less often to aviation documents but could assist with knowledge management and readability of documents [32]. Ultimately, NLP techniques show promise for augmenting tasks traditional performed by analysts, such as report classification, as well as improving knowledge management as documents increasingly become digitally available. Because aviation has highly specialized vocabulary [20] and domain-specific BERT models have improved performance on specialized tasks, the proposed SafeAeroBERT model aims to provide a language model to improve aviation NLP results.

IV. Methodology

To build a safety-informed aviation specific language model, a pre-existing BERT model is pre-trained on aviation safety reports, then fine-tuned for specific tasks in accordance with Figure 1. In this section, we first describe the training data used from the Aviation Safety Reporting System (ASRS). Next, we explain the architecture of a BERT

language model, followed by the method we use to pre-train the BERT model on ASRS data. Procedures for fine-tuning the model on different tasks (i.e., classification, NER, RE, IR, QA, and summarization) are discussed in detail.

Figure 1: Pre-training and fine-tuning set up for SafeAeroBERT.

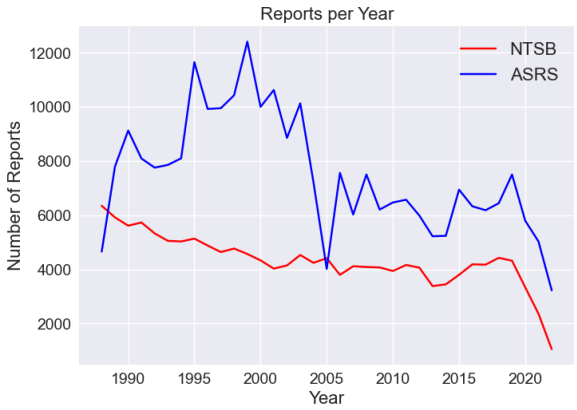


A. Data

The Aviation Safety Reporting System (ASRS) is the primary dataset used in this research and is a confidential and voluntary reporting system founded in 1976 that can be used by any operator or personnel participating in aviation operations, including pilots and air traffic controllers [33]. Reports should be filed in response to any incident where safety is or has the potential to be jeopardized [33]. Due to the voluntary nature of the ASRS, there is inevitable self-reporting bias which may exhibit through inaccurate descriptions of events, missing reports entirely, etc. Personnel are however strongly encouraged to always file ASRS reports when deemed necessary as reports undergo a thorough deidentification process and provide limited immunity for reporting parties [33]. Upon the submission of a report, at minimum two highly experienced analysts review the report to identify any reported hazards, notify any necessary organizations (e.g., the FAA), and identify causes of any reported hazards [33]. Because ASRS reports contain many abbreviations, abbreviations are expanded to full words and reports are cleaned using standard procedures prior to use.

The National Transportation Safety Board (NTSB) also maintains a database of aviation accident reports and specializes in causes of civil aviation accidents [34]. Reports are completed according to an investigative process by the NTSB in response to accidents of interest [34]. After an accident occurs, the NTSB decides to launch an investigation or not [34]. Once the investigation begins, NTSB personnel analysis data leading up to the accident and from the accident site [34]. Data are then analyzed to determine the sequence of events leading to the accident, with an emphasis on the probable cause [34]. A final report is created after the investigation, and safety recommendations may be provided to prevent similar accidents [34]. Due to the rich narrative components and safety emphasis, NTSB reports containing are used for pre-training SafeAeroBERT in addition to ASRS reports. In total, SafeAeroBERT is pre-trained on 161,460 NTSB reports and 262,935 ASRS reports, where each report may contain multiple narrative sections. As seen in Figure 2, ASRS reports outnumber NTSB reports in almost every year, with the frequency of reports steadily decreasing over time.

Figure 2: Number of ASRS and NTSB reports in the pre-training dataset from 1990 to 2022.



B. BERT Model

Bidirectional Encoder Representations from Transformers (BERT) models use an architecture with encoders and attention mechanisms and train using a masked language model. Hence, specified words will be masked or removed from a sentence during training to learn the context of the word. By default, 15% of word tokens are masked during training. BERT also trains on next sentence prediction. That is, given a pair of sentences, if the two sentences are related (i.e., one sentence follows the other), then the model learns a higher probability than if the two sentences are unrelated. Through next sentence prediction, BERT learns the context of sentences within a text, whereas through masking BERT learns the context of words within a sentence. However, recent research developing the roBERTa variant has demonstrated next sentence prediction is unnecessary for training a BERT model and can decrease performance on downstream tasks [13]. The encoder components essentially decrease the dimensionality of the data, while the attention mechanism allows the model to consider all words in the sentence and weigh them according to importance.

The base BERT model was trained on approximately 16 GB of data from the English Wikipedia and a corpus of books for one million steps with a batch size of 256 documents [1]. Prior to inputting data to the model, BERT uses a tokenizer to break sentences into a sequence of sub-word pieces defined by the vocabulary, which consists of 30,000 word pieces in the base model [1]. The base model has 12 layers with 768 neurons and 12 attention heads, resulting in 110 million parameters in the model [1]. The large variant instead has 24 layers with 1024 neurons and 16 attention heads, with a total of 340 million parameters [1].

C Pre-training

SafeAeroBERT is pretrained only on masked language modeling in accordance with the original BERT model and findings from the roBERTa variant [13]. We train the model on masked language modeling using the Huggingface Transformers [35] python package with a pytorch configuration and use the default loss function, which is cross entropy loss. The final model was trained for just over two epochs across three NVIDIA Tesla GPUs with a training batch size of eight and sixteen gradient accumulation steps, resulting in a total batch size of 384. This resulted in 13,000 training steps, which took approximately two months with occasional downtime. A total of 2,283,435 narrative sections and 1,169,118,720 tokens from over 400,000 NTSB and ASRS documents are used for model pre-training.

D. Fine-tuning: Classification

To evaluate the performance of SafeAeroBERT, the model is fine-tuned on an upstream task of document classification; however, future work should evaluate the model on named-entity recognition, information retrieval, question answering, and summarization as well. SafeAeroBERT is fine tuned for document classification using a supervised learning approach. ASRS documents come prelabeled with contributing factors, which is the primary field for classifying and categorizing reports. Each report has one or more contributing factor and example contributing factors include weather, human factors, procedures, and aircraft. Thus, assigning contributing factors to reports is a multilabel classification problem. For fine-tuning SafeAeroBERT, we transform the problem using binary relevance which instead builds one binary classifier for each contributing factor rather than one large multilabel classifier. For each contributing factor, an extra dense layer of one neuron is added onto SafeAeroBERT for document classification. Both the base BERT model and SciBERT are also fine-tuned using the same method as a comparison. The training, test, and validation set consists of three contributing factors, with document frequencies listed in Table **2**. Documents not containing the specified contribution factor are randomly selected as well at the same frequency. That is for contributing factor ‘Procedure’, we have 1,172 documents with the contributing factor and 6,828 documents without the contributing factor in the training set. Each model is fine-tuned with a binary cross-entropy loss and learning rate of 0.00002 for five epochs with batch size of thirty-two across three NVIDIA Tesla GPUs.

Table 2: Frequency of documents for each contributing factor in training, validation, and test sets.

Contributing Factor		Train	Validation	Test
Aircraft	Present	2083	260	260
	Not Present	5917	740	740
	Total	8000	1000	1000
Human Factors	Present	3609	451	451
	Not Present	4391	549	549
	Total	8000	1000	1000
Procedure	Present	1172	146	146
	Not Present	6828	854	854
	Total	8000	1000	1000
Weather	Present	1035	129	129
	Not Present	6965	871	871
	Total	8000	1000	1000

V. Results

A. Classification

Document classification results are shown in Table 3, with evaluation metrics accuracy, precision, recall, and f-1 displayed. Accuracy is the total percent of classifications that are correct, while precision is the percent of positive classifications that are true positives and recall is the percent of true positives that are correctly identified by the model. F-1 measures an average of precision and recall. From the results, SafeAeroBERT performs better than the other models on classifying accidents due to procedural errors and weather, while BERT and SciBERT perform better on human factors and aircraft driven reports. SafeAeroBERT tends to have a higher accuracy and recall, indicating that it has better success at identifying all of the positive cases, potentially at the cost of a lower precision and higher false positive rate. While these results show a mixed performance of SafeAeroBERT in comparison to the other models, early results where models were trained under the same parameters on CPU and with raw ASRS reports found SafeAeroBERT performing consistently better on all classification tasks. In this initial study, all models were trained under the same parameters; however, the optimal classification for each model may need to be obtained with specialized training parameters for each model and task pair. Ultimately these results do indicate increases in accuracy in some cases when using a specialized pre-trained model in comparison to the default general BERT model.

Table 3: Classification metrics for SafeAeroBERT vs base BERT and SciBERT on the test set, with the best score in each row bolded.

Contributing Factor	Metric	BERT	SciBERT	SafeAeroBERT
Aircraft	Accuracy	**0.747**	0.726	0.740
	Precision	**0.716**	0.691	0.548
	Recall	**0.747**	0.726	0.740
	F-1	**0.719**	0.699	0.629
Human Factors	Accuracy	**0.608**	0.557	0.549
	Precision	**0.618**	0.586	0.527
	Recall	**0.608**	0.557	0.549
	F-1	**0.572**	0.426	0.400
Procedure	Accuracy	0.766	0.755	**0.845**
	Precision	**0.766**	0.762	0.742
	Recall	0.766	0.755	**0.845**
	F-1	0.766	0.758	**0.784**
Weather	Accuracy	0.807	0.808	**0.871**
	Precision	**0.803**	0.769	0.759
	Recall	0.807	0.808	**0.871**
	F-1	0.805	0.788	**0.811**

VI. Discussion

Despite having access to powerful computational resources, the time required to train SafeAeroBERT was prohibitive at over two months of run time. However, pre-training should be a one-time activity that allows others to further fine-tune the model for specific tasks. In the classification tasks SafeAeroBERT was fine-tuned for, SafeAeroBERT performed as good or better than the comparison models of BERT and SciBERT. Both BERT and SciBERT are intended to be general use models, with SciBERT further trained for scientific documents, yet these results indicate that these models are not as effective with aviation documents as the domain-specifc SafeAeroBERT model. With this in mind, other specialized aviation NLP tasks, such as named-entity recognition for FMEAs, could be improved by fine-tuning SafeAeroBERT over standard BERT.

A complete large language model specific to aviation safety requires high performance across the set of relevant aviation natural language processing tasks. While this research presents a pre-trained model, this model's effectiveness on other aviation natural language processing tasks must still be evaluated. However, all domain specific language models should undergo performance validation across a range of tasks, so this limitation is not unique to SafeAeroBERT. The computation resources and time needed to train SafeAeroBERT are not accessible to all researchers, but the model itself should be available for others to fine-tune and evaluate in the future. Since SafeAeroBERT only slightly outperforms the base BERT model, further evaluation should examine if the performance gains are tangible and compensate for the computational cost.

VII. Conclusion

This research presents progress towards an aviation-specific safety informed language model by pre-training a BERT model on aviation safety reports and evaluating the model on a specific document classification task. The proposed SafeAeroBERT model shows promise for improved performance in natural language processing tasks when compared to the base BERT and SciBERT, as evident by the classification task results. Future work includes evaluating the model on other natural language tasks, including question answering, named-entity recognition, and information retrieval. Natural language processing techniques have advanced in recent years, with numerous domain-specific models exhibiting improved performance over general use case models. Aviation generates large numbers of text reports, yet the information contained in these repositories may not be fully utilized with general use language models. Hence, the SafeAeroBERT prototype provides a safety-centric aviation language model that could improve performance on various natural language processing tasks currently performed in aviation research and operations.

Acknowledgments

This research is supported by the System-Wide Safety (SWS) project in the Airspace Operations & Safety Program (AOSP) in the NASA Aeronautics Research & Mission Directorate (ARMD) funded Under Prime Contract No. 80ARC020D0010 with the NASA Ames Research Center. Any opinions or findings of this work are the responsibility of the authors and do not necessarily reflect the views of the sponsors or collaborators.

References

[1] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv*, 2018.

[2] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, 2019.

[3] D. Khurana, A. Koli, K. Khatter and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, 2022.

[4] D. W. Otter, J. Medina, J. Medina and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-21, 2020.

[5] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3-26, 2007.

[6] N. Bach and S. Badaskar, "A review of relation extraction," *Literature review for Language and Statistics II*, vol. 2, pp. 1-15, 2007.

[7] W. Ali, W. Zuo, R. Ali, X. Zuo and G. Rahman, "Causality Mining in Natural Languages Using Machine and Deep Learning Techniques: A Survey," *Applied Sciences*, 2021.

[8] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences,* vol. 63, no. 10, pp. 1872-1897, 2020.

[9] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds and Machines,* vol. 30, no. 4, 2020.

[10] K. Ethayarajh, "How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings," *arXiv,* 2019.

[11] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084,* 2019.

[12] V. Sanh, L. Debut, J. Chaumond and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv,* 2019.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv,* 2019.

[14] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[15] D. T. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained," *arXiv,* 2019.

[16] A. Brandsen, S. Verberne, K. Lambers and M. Wansleeben, "Can BERT Dig It? – Named Entity Recognition for Information Retrieval in the Archaeology Domain," *Journal on Computing and Cultural Heritage,* 2022.

[17] I. Beltagy, K. Lo and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, 2019.

[18] M. Alomari, *Introducing AeroBERT, An Aerospace-centric Language Model for all NLP Applications at Rolls-Royce,* NVIDIA, 2021.

[19] A. Berquand, P. Darm and A. Riccardi, "SpaceTransformers: Language Modeling for Space Systems," *IEEE Access,* vol. 9, pp. 133111-133122, 2021.

[20] N. Amin, T. L. Yother, M. E. Johnson and J. Rayz, "Exploration of Natural Language Processing (NLP) Applications in Aviation," *Collegiate Aviation Review International,* vol. 40, no. 1, pp. 203-216, 2022.

[21] T. Dong, Q. Yang, N. Ebadi, X. R. Luo and P. Rad, "Identifying Incident Causal Factors to Improve Aviation Transportation Safety: Proposing a Deep Learning Approach," *Journal of Advanced Transportation,* 2021.

[22] X. Zhang and S. Mahadevan, "A Hybrid Data-Driven Approach to Analyze Aviation Incident Reports," in *AIAA Aviation Forum*, 2018.

[23] R. L. Rose, T. G. Puranik and D. N. Mavris, "Natural Language Processing Based Method for Clustering and Analysis of Aviation Safety Narratives," *Aerospace,* vol. 7, no. 10, 2020.

[24] C. Paradis, R. Kazman, M. Davies and B. Hooey, "Augmenting Topic Finding in the NASA Aviation Safety Reporting System using Topic Modeling," in *AIAA Scitech Forum,* 2021.

[25] S. Robinson, "Temporal topic modeling applied to aviation safety reports: A subject matter expert review," *Safety Science,* vol. 116, no. 2, pp. 275-286, 2019.

[26] R. L. Rosea, T. G. Puranika, D. N. Mavrisa and A. H.Rao, "Application of structural topic modeling to aviation safety data," *Reliability Engineering and System Safety,* vol. 224, 2022.

[27] S. R. Andrade and H. S. Walsh, "Machine Learning Enabled Quantitative Risk Assessment of Aerial Wildfire Response," in *AIAA Aviation Forum*, 2022.

[28] R. Pai, S. S. Clarke, K. Kalyanam and Z. Zhu, "Deep Learning based Modeling and Inference for Extracting Airspace Constraints for Planning," in *AIAA Aviation Forum*, 2022.

[29] A. T. Ray, O. J. Pinon-Fischer, D. N. Mavris, R. T. White and B. F. Cole, "aeroBERT-NER: Named-Entity Recognition for Aerospace Requirements Engineering using BERT," in *AIAA Scitech*, 2023.

[30] S. R. Andrade and H. S. Walsh, "What Went Wrong: A Survey of Wildfire UAS Mishaps through Named Entity Recognition," in *2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC)*, 2022.

[31] S. Kierszbaum and L. Lapasset, "Applying Distilled BERT for Question Answering on ASRS Reports," in *2020 New Trends in Civil Aviation (NTCA)*, 2020.

[32] A. V. Martin and D. Selva, "Explanation Approaches for the Daphne Virtual Assistant," in *AIAA Scitech Forum*, 2020.

[33] B. L. Hooey, "ASRS Program Briefing," NASA, 2021.

[34] National Transportation Safety Board, "The Investigative Process," 2022. [Online]. Available: https://www.ntsb.gov/investigations/process/Pages/default.aspx. [Accessed 2022].

[35] T. Wolf and e. al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.