EPiC Series in Computing

Volume 102, 2024, Pages 189–198

Proceedings of Conference on Cognitive and Computational Aspects of Situation Management 2023

Building Holistic Situation Awareness through Large Language Models *

Emma L. McDaniel[1] and Alicia I. Ruvinsky[2]

[1] Georgia State University, Atlanta, GA, USA
Department of Computer Science
emcdaniel10@gsu.edu

[2] US Army Engineer Research and Development Center, Vicksburg, MS, USA
Cybersecurity Engineering and Analysis Branch
alicia.i.ruvinsky@erdc.dren.mil

Abstract

In order to design a system to support a users' situation awareness as they navigate and execute in a complex domain, the information landscape for the domain must be defined. Such a landscape becomes critical for recognizing and evaluating deficiencies in a user's cognitive processing of the information landscape, and as such, the user's state of situation awareness. This paper leverages previous work in defining the information landscape for the domain of rotorcraft pilotage, and explores an approach to identifying common deficiencies in a user's awareness of the information landscape. This paper takes into account prominent frameworks on situation awareness, and presents a methodology leveraging Large Language Models (LLMs) to identify common information deficiencies occurring at various phases of rotorcraft flight through text mining aviation incident and accident reports. Once these deficiencies are identified, the system can provide holistic situation awareness considering a larger data space than the user, and also provide human-centered awareness of prioritized concerns of situation awareness factors and possible mitigations.

1 Introduction

Situation awareness is the state of knowing what is happening in one's environment, and projecting the implications of these current happenings onto the present as well as into the future [1]. A significant aspect of making better decisions in a cognitively taxing context, such as disaster response, hostile engagements, or degraded environments, is to maintain situation awareness despite characteristics of such contexts like compromised sensors or disorienting information. In order to design a system to support the persistence of a users' situation awareness as they

*This research was supported in part by an appointment to the Department of Defense (DOD) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the DOD. ORISE is managed by ORAU under DOE contract number DE-SC0014664. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of DOD, DOE, or ORAU/ORISE.

K. Baclawski, M. Kozak, K. Bellman, G. D'Aniello, A. Ruvinsky and C. Da Silva Ferreira Barreto (eds.), CogSIMA 2023 (EPiC Series in Computing, vol. 102), pp. 189–198

navigate and execute in a complex domain, the system must capture and manage an awareness space larger than that of the user. In other words, the system must construct and maintain a holistic situation awareness of the domain, and persistently assess the user's specific situation awareness within it. An approach to doing this is to build an information landscape for the domain. Such a landscape becomes critical for recognizing and evaluating deficiencies between the user's current state of situation awareness and the user's needed state. For example, in designing an information landscape for aircraft flight, it is clear that terrain awareness during landing is critical. As such, if a user is landing an aircraft during a sandstorm, the system would ensure that critical terrain information such as power lines, buildings, or mountains are presented to the user.

This paper leverages previous work in defining the information landscape for the domain of rotorcraft pilotage, and explores a methodological approach to identifying common deficiencies in a user's awareness of the landscape. This paper takes into account prominent frameworks on situation awareness, and presents a methodology leveraging Large Language Models (LLMs) to identify common information deficiencies occurring at various phases of rotorcraft flight through text mining aviation incident and accident reports. Once these deficiencies are identified, the system can provide a holistic situation awareness considering a larger data space than the user, and also provide a human-centered awareness of prioritized concerns of situation awareness factors and possible mitigations.

The proposed experimental design uses a fine-tuned LLM to mine historical helicopter aviation incident and accident narratives. The goal of the mining is to identify information deficiencies that may occur because of cognitive overload or if situational awareness is otherwise compromised. Our approach involves generating various sets of sentences that describe possible problems related to functional components of a helicopter flight. Then, we embed these generated sets of sentences and the sentences from the aviation incident and accident narratives using the fine-tuned LLM. Once embedded, we compare the similarity between the generated sets of sentences to that of narrative sentences. For each narrative sentence, if it is close enough in semantic distance either to a set of generated sentences, it will be labeled with the information deficiency to that functional component. For instance, if a sentence from a narrative is close in semantic space to that of sentences describing a fuel-related information deficiency, the narrative sentence will be labeled as a fuel-related information deficiency. This process enables us to prioritize information deficiency concerns in relation to a pilot's situation awareness based on historical data and by state of helicopter.

Challenges emerge when using LLMs as a tool within our proposed methodology. A particular challenge related to utilizing embeddings created with a LLM is the risk of not being able to discriminate between minimal but crucial differences in sentences. For instance, two sentences describing an unexpected landing around power lines where one landing is at night and the other is during the day; the LLM is less likely to be able to discriminate between the two sentences without domain-specific fine-tuning. Another challenge is regarding the use of generative LLMs; these involve the potential for hallucinations, where the model's output may not accurately align with the given input. The potential for hallucinations as they relate to our intended usage could result in three possibilities for the generated sentences: 1. the narrative sentences will not be close enough in semantic distance to the hallucinated sentences rendering them not usable for labeling and thus self-pruning; 2. the narrative sentences could be close in semantic distance to the hallucinated sentences and thus prove valuable in understanding the underlying information deficiency; 3. and, lastly, the narrative sentences will be semantically close to the hallucinated sentences, but lack coherence, necessitating mitigation of this issue. As our research unfolds and instances of these possibilities emerge, our understanding will improve in order to consider and design mitigations. We hope to tackle this phenomenon in future work.

The goal of this work is to build an information deficiency landscape to compliment the information landscape defined in previous work [2]. The information deficiency landscape will identify the kind of information that, if missing from a pilot's situation awareness, may lead to incident or accident. In evaluating this deficiency landscape in conjunction with the information landscape, the aim of the research is to ultimately predict informational needs, preempt the pilot's awareness of this information, prevent loss of situation awareness, and thereby improve the persistence of situation awareness throughout flight. This paper is a presentation of our approach, experimental design, and progress to date. First, the paper describes a brief background of the open source data leveraged for this work as well as LLMs and what tools will be used for this work. This is followed by a *Methodology* section including a description of the proposed experimental design. The state of this work to date with *Preliminary Work* is then followed by *Future Work* needed to bring this work to fruition. Lastly, in the *Conclusion*, we resituate the proposed methodology within a larger decision support system framework.

2 Background

Datasets

For the proposed methodology, we intend to utilize two open source datasets of aviation accident and incident records in the United States; these two datasets are curated by: the National Transportation Safety Board (NTSB) [3] and National Aeronautics and Space Administration's (NASA) Aviation Safety Reporting System (ASRS) [4]. The NTSB dataset has 6,111 helicopter records and the NASA-ASRS has 2,434 helicopter records. These records implicitly involve a negative occurrence, which we will leverage when identifying information deficiencies of pilots. A limitation of these datasets, is that they may not be able to provide insights on critical tasks that are typically executed, but bear significant consequences if not performed. This is not an exhaustive experiment, but instead a proof of concept for utilizing LLMs to identify components of the information landscape as part of a robust, ensembled model with complementary capabilities designed to target distinct aspects of the landscape.

Large Language Models

The prevailing architecture for LLMs builds upon the Transformer architecture originally proposed by Vaswani et al. (2017)[5]. Transformers are a type of encoder/decoder neural network that are specifically designed for sequence modeling. The encoder portion of the Transformer model outputs embeddings that can represent the semantic meaning of the inputted natural language in a numerical format. We used three models that utilize only the encoder portion of the Transformer for our preliminary results; however due to space constraints, we only report our findings using “all-mpnet-base-v2” [6]. The two other models “all-distilroberta-v1” [7] and “all-MiniLM-L6-v2” [8] are included in our online project repository https://osf.io/ebv48/. The base model of “all-mpnet-base-v2” is the MPNet model [9]. MPNet is trained using two different strategies Masked Language Modeling (like BERT [10]) Permuted Language Modeling (like XLNet [11]). This model was fine-tuned for semantic similarity following the method as described in Reimers and Gurevych (2019) [12] where it received high scores compared to other models on a variety of tasks related to semantic similarity [13]. This model was then further fine-tuned as described on the HuggingFace model card [6] and has a high number of downloads for models dealing with the task of semantic similarity on HuggingFace.

Aviation Narratives and Natural Language Processing

Unstructured data, like narratives, contain crucial information about an event that oftentimes does not get transferred into a structured format. When analyzing records, important components of the event can be missing because of the lack of context or ability to transfer all information into a given data schema. Also, these records may be created for a specific purpose (e.g. recommendation for aviation safety) but can be mined for another purpose (e.g. landscape of information deficiencies). The methodology of utilizing word/sentence embeddings and/or language models like LLMs is becoming increasingly common for extracting information. In the domain of natural language processing and aviation reports, there is much research on topic modeling, unsupervised classification of features, and identifying broad patterns. Across other domains, text mining narratives have been used to identify patterns that otherwise would not have been as easily extracted.

To better understand patterns from aviation reports, there have been many efforts to first embed a narrative (not using LLMs) and then employ differing unsupervised techniques to classify the narratives. Both Rose et al. (2020)[14] and Miyamoto et al. (2022) [15] utilize TF-IDF matrices as their embeddings which they then cluster; Rose et al. [14] attempted to identify trends among narrative categories and Miyamoto et al. [15] worked to identify inefficient operational patterns. In both Chanen's (2016) [16] and Seale et al.'s (2019) [17] work, they trained word2vec on domain specific aviation narratives and were able to identify terms that were used in semantically similar manner to a subject matter expert list so to supplement information retrieval and data analytics. Luo and Shi (2019) [18] devised a method to use a word2vec and Latent Dirichlet Allocation (LDA) modeling to identify topics of aviation safety reports. Zhang et al. (2021) [19] use a Long, Short-Term Memory network with inputs from a Keras tokenizer (similar to a 1-hot encoding scheme but preserves the order of words) to identify if a report is an incident or accident. Madeira et al. (2021) [20] attempted to classify human factor categories from aviation reports using a doc2vec strategy with Bayesian Optimization for hyperparameter choice.

More recently, aviation narrative mining efforts are using Transformer-like models to identify broad patterns in the narratives as it relates to the researchers' goals. Kierszbaum et al. (2022) [21] tested whether or not a trained from scratch using low-volume in domain data from NASA ASRS (ASRS-CMFS) or a base pre-trained RoBERTa model would have better performance on an in-domain natural language understanding task. They found that the two models performed similarly but with RoBERTa at times achieving higher scores. They also found that the smaller model, ASRS-CMFS was more efficient because of its training time and size, but also did not perform much worse than the large RoBERTa. Chandra et al. (2023) [22] created a Aviation-Specific Tokenizer and also fine-tuned the BERT base, which they call Aviation-BERT. Their model achieves better results at producing the correct term within its top five results in a text masking problem than that of the original BERT and a secondary model like Aviation-BERT but without their custom tokenizer. In Jonk et al. (2023) [23], they report three research efforts using natural language processing on aviation reports, two of which utilize a Transformer; the first Transformer project utilizes a BERT-based classifier trained on NASA's ASRS and Canada's Civil Aviation Daily Occurrence Reporting System to label records into an occurrence topic for NTSB data; and, the other Transformer project performed automatic probable cause summary generation using a BART-based model trained on NTSB reports. Further research still needs to be conducted on the mining of specific information from unstructured data using LLMs.

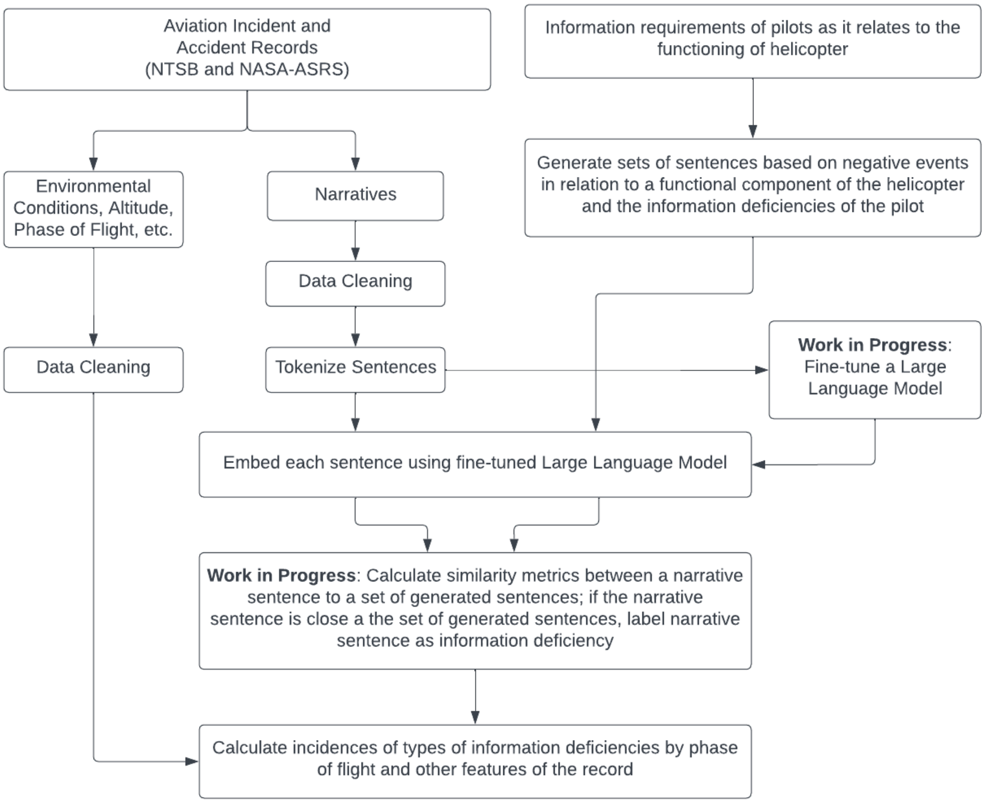
Aviation Incident and Accident Records (NTSB and NASA-ASRS)
Information requirements of pilots as it relates to the functioning of helicopter
Environmental Conditions, Altitude, Phase of Flight, etc.
Narratives
Generate sets of sentences based on negative events in relation to a functional component of the helicopter and the information deficiencies of the pilot
Data Cleaning
Data Cleaning
Tokenize Sentences
Work in Progress: Fine-tune a Large Language Model
Embed each sentence using fine-tuned Large Language Model
Work in Progress: Calculate similarity metrics between a narrative sentence to a set of generated sentences; if the narrative sentence is close a the set of generated sentences, label narrative sentence as information deficiency
Calculate incidences of types of information deficiencies by phase of flight and other features of the record

Figure 1: Experimental Design for Mining Aviation Incident/Accident Narratives

3 Methodology

Large Language Models (LLMs) excel at handling words with similar meanings (synonymy) and have been shown to manage the contextual variations of words within a given sentence. This proposed methodology is a work in progress and is driven by evidential support and a literature review that has been, in part, already completed. At a high-level, the goal of this research is to label the information deficiencies that occurred during a rotorcraft flight leading to an incident/accident event by comparing sentences from the aviation incident and accident report narratives to pre-generated sentences that are of a known specific category of information deficiency. The information needs are devised from subject matter interviews with pilots about their self-perceived information needs by phase of flight [2]. The prompts used to create the generated sentences come from these interviews and an example prompt is provided in *Preliminary Work*. If a similarity measure between a narrative sentence is close to that of a set of generated sentences, then the narrative sentence will be labeled as describing that information deficiency.

This experimental design consists of two main branches, as depicted in Figure 1; and this is still a work in progress. The left branch begins with the collection of aviation incident and accident records from NTSB and NASA-ASRS. Subsequently, the collected data undergoes cleaning and tokenization. The sentences extracted from these narratives are then used to fine-tune a LLM to enhance its performance as it relates to semantic similarity in the aviation domain. The right branch of Figure 1 begins with preliminary work done by a project team at U.S. Army Engineer Research and Development Center (ERDC) where they completed subject matter interviews with pilots about their self-perceived information needs by phase of flight [2]. From these information needs, we devise a list of potential information deficiencies, and prompt a generative LLM to create sentences about these information deficiencies as it relates to helicopter functionality. The generated sentences and each narrative sentence will be embedded using the fine-tuned LLM. The generated and narrative sentence embeddings will then be compared using a similarity metric and if close enough to a set of generated sentence embeddings then it will obtain the label of that information deficiency. Close enough is defined as either utilizing a distance threshold or a clustering methodology. In the final step, the labeled sentences will be counted by differing states of the rotorcraft, such as flight phase or the environment (e.g night/day). These incidence counts will provide insights into common information deficiencies by a state of the rotorcraft. It can specifically assist us in identifying the information that tends to be overlooked during periods of cognitive overload for pilots.

4 Preliminary Work

To assess whether a generic LLM (non-aviation domain) fine-tuned for semantic similarity would be sufficient for use within our research, we conducted preliminary experimentation. The code for data cleaning and implementation of the following experiment can be found at: https://osf.io/ebv48/. The design of this preliminary experiment compares the cosine distances between three sets of sentence embeddings in order to assess whether or not the proximity strategy for labeling will be reliable.

These three groups of sentences are: a set of generated sentences about a specific functional component of the helicopter, a sentence from a narrative that is about the specific functional component of the helicopter, and unrelated sentences. See Table 1 for the specific sentences as well as the labels for the embeddings in the distance matrix Table 2. To create the generated sentences (0g-9g) about a specific functional component of the helicopter, we prompted ChatGPT 3.5 [24] to generate 10 sentences related to fuel problems. The prompt utilized was:

```
Generate 10 sentences describing negative issues on a helicopter related to fuel,
crossfeed, fuel level, tank, fuel burn, flamed out, and flame out.
```

The narrative sentence [10ts] describes an event leading to when a helicopter flames out. The three unrelated sentences [11n-13n] are: about a boat having a fuel related problem, problem with pizza at a restaurant, and a helicopter having a near miss. The last embedding [14a] is an average of the first 10 generated sentences [0g-9g]. The average was computed in order to identify if this could be a possible technique to identify related narrative sentences to the specific information deficiency (e.g. fuel) more efficiently.

The sentences were embedded using three different models fine-tuned for semantic similarity, and only one is included in this results section for reasons described in *Background: Large Language Model*. We calculate the cosine distance between each of the embeddings for the 10 generated sentences, a narrative sentence, three unrelated sentences, and the average of the generated sentences. Cosine distance is computed by obtaining the difference between 1 and the cosine similarity of two embeddings/vectors. Cosine similarity is calculated by computing the dot product of the two vectors (A and B) divided by the product of the magnitude of the two vectors.

$$CosineDistance = 1 - \frac{A \cdot B}{||A|| * ||B||}$$

Table 1: Sentences or description of the 15 embeddings in Table 2 distance matrix

Table 2 Tag	**Sentences/Description**
0g	The helicopter experienced a fuel crossfeed malfunction, leading to an imbalance in fuel distribution and potential stability issues.
1g	A sudden drop in fuel level during the flight raised concerns about possible fuel leakage or inefficient fuel management.
2g	The fuel tank developed a significant leak, posing a serious safety hazard and necessitating immediate maintenance.
3g	Unexpectedly high fuel burn rate indicated a fuel efficiency problem, requiring investigation and corrective measures.
4g	Both engines flamed out due to a complete fuel depletion, resulting in an emergency autorotation landing.
5g	Fuel contamination in the helicopter's tank compromised the fuel quality, causing engine performance degradation.
6g	Inadequate fuel monitoring and planning led to an unforeseen fuel shortage during a critical helicopter operation.
7g	The fuel gauge malfunctioned, providing inaccurate readings and impeding accurate assessment of the fuel level.
8g	Excessive fuel burn during hover operations drained the fuel reserves faster than anticipated, creating a hazardous situation.
9g	The helicopter experienced a flameout during a steep descent due to insufficient fuel supply, necessitating an emergency landing.
10ts	The pilot further reported that during the approach to land he made a right turn to final and the engine 'flamed out'.
11n	The boat had a fuel problem.
12n	My pizza at a restaurant this weekend was cold, which resulted in my attempt to get a refund.
13n	The helicopter was flying too high and was in the incorrect airspace, which caused a near miss with a personal aircraft.
14a	Average of generated fuel sentences [0g-9g]

Table 2: Cosine distance matrix between the embeddings obtained using the MPNet model [6]

	1g	2g	3g	4g	5g	6g	7g	8g	9g	10ts	11n	12n	13n	14a
0g	0.36	0.45	0.48	0.46	0.24	0.28	0.37	0.37	0.38	0.49	0.54	0.88	0.42	0.18
1g		0.34	0.37	0.51	0.32	0.32	0.29	0.31	0.42	0.46	0.47	0.84	0.59	0.16
2g			0.47	0.50	0.35	0.46	0.42	0.38	0.49	0.46	0.47	0.91	0.67	0.23
3g				0.52	0.36	0.44	0.41	0.34	0.52	0.47	0.52	0.85	0.70	0.24
4g					0.46	0.52	0.50	0.37	0.36	0.28	0.59	0.89	0.49	0.28
5g						0.32	0.35	0.32	0.40	0.47	0.43	0.89	0.46	0.14
6g							0.40	0.36	0.34	0.53	0.52	0.83	0.49	0.18
7g								0.42	0.52	0.46	0.45	0.86	0.66	0.21
8g									0.35	0.36	0.49	0.86	0.51	0.15
9g										0.33	0.54	0.86	0.41	0.22
10ts											0.52	0.86	0.52	0.29
11n												0.90	0.77	0.38
12n													0.89	0.83
13n														0.42

The distance matrix in Table 2 reports the cosine distance between all the embeddings which ultimately summarizes the proximity relationship between the generated sentences [0g-9g], the not related sentences [11n-13n], and the narrative sentence[10ts] as produced by the LLM. The duplicate rows and columns of the distance matrix are not included in Table 2.

We selected a threshold of 0.3 distance to gauge proximity and examined whether the generated sentences, their average, and the narrative sentence met this threshold. This threshold was chosen instead of a clustering method because of an inability to perform due to the small number of samples used for this preliminary experiment.

The 10 generated sentences [0g-9g] in Table 2 exhibit varying degrees of closeness; in general, they are not within the 0.3 threshold distance to each other despite all the sentences being about a fuel-related issue on a helicopter. The sentence, 10ts, from an aviation incident/report narrative, was barely within the distance threshold to the generated sentence. This closest generated sentence, 4g, at 0.28 specifically referred to the fuel-related problem as flame-out similar to that of the narrative sentence. The unrelated sentences, 11n-13n, were not within the distance threshold to the narrative and generated sentences. However, 13n, a sentence describing a helicopter having other issues not related to fuel, was at times closer to the generated fuel sentences than the narrative sentence. The average of the embeddings of the generated sentences, 14a, is a 0.29 distance from the narrative sentence, which would be within the 0.3

threshold, but we had anticipated that it would be sufficiently closer.

In order to more confidently identify information deficiencies described in incident and accident report narratives, the distances between the embeddings of the narrative sentences and generated sentences of a known information deficiency need to be lessened. Through further fine-tuning of a LLM using aviation domain specific datasets, the model may more adeptly encode these nuances into its embeddings. Thus facilitating the capability to label narrative sentences based on their proximity to a set of generated sentences of a known information deficiency.

5 Future Work

This project is a work in progress and as such future work encompasses much of what this project entails. There are already exciting opportunities for further research and analysis related to this work that may be done in conjunction with the proposed methodology. One compelling direction is the expansion of datasets by including other countries' aviation accident and incident records like the European Coordination Centre for Accident and Incident Reporting Systems or Canada's Civil Aviation Daily Occurrence Reporting System. This could benefit the fine-tuning of the LLM and also provide more data for the incidence counts of information deficiencies in the rotorcraft pilot information landscape. Another potential research direction is exploring the adaptability of this methodology to other domains. One such application could involve utilizing storm event narratives from the National Oceanic Atmospheric Association (NOAA) storm events database. This methodology when applied to this data would instead identify damage/impacts on critical infrastructure caused by the storm events. This would then provide the possibility for disaster preparedness practitioners to make data-driven decisions that can improve planning efforts for building disaster resilience.

6 Conclusion

Situation awareness is critical component of decision making. Designing mechanisms for computation support to decision makers by enabling, augmenting, and persisting situation awareness as environment, mission, or other context changes is an impactful and important area of research. A significant challenge to building computational support for situation awareness within a decision support system consists of representing the situational state of the world as well as the user's awareness state within that situation.

Overcoming this representational barrier will require various tools and capabilities such as world modeling and change monitoring, but the capability being investigated here regards identifying information needs of the user based on historic information deficiencies within a certain context. For example, recognizing a context of interest such as a particular phase of flight (e.g., take-off or descent), and recognizing historic information deficiencies that have occurred within this context. Such an association will drive the prescribed information dissemination to a user encountering the context of interest. In other words, considering incidents and accidents that have occurred in the past due to a user's lack of awareness to specific information, the system aims to identify this information deficiency and mitigate the loss of situation awareness to future users in similar context.

The work presented here is targeting the semi-supervised labeling of human-generated narrative text. Previous work [2] captured a subject matter expert-based decomposition of information requirements at each phase of rotorcraft flight. Given this kind of decomposition of informational needs for performing a task or set of tasks within a domain, this work hypothesizes that given such an ontology of information need, one can generate a set of contrapositive assertions of effects resulting from deficiency of that information. Of particular value is that these generated assertions will also be labeled with the information deficiency that it is depicting. The work further hypothesizes that with these labeled contrapositive assertions, the labeling of human-generated narrative text may be automated based on the relative semantic similarity of the narrative text to the labeled assertions. Though this is a work in progress, there appears to be promise in the methodology and in the specific approach designed for the rotorcraft pilotage domain.

References

[1] M. Endsley, “Situation awareness global assessment technique (sagat),” in *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, pp. 789–795 vol.3, 1988.

[2] R. C. Salter, M. A. Clement, and A. I. Ruvinsky, “INTERNAL: An Overview of Data and Information Necessary for Successful Rotorcraft Pilotage,” tech. rep., Engineer Research and Development Center, 2023.

[3] “National Transportation Safety Board (NTSB).” Date Downloaded: June 20, 2023, Online: https://data.ntsb.gov/avdata.

[4] “NASA Aviation Safety Reporting System (ASRS).” Date Downloaded: June 26, 2023, Online: https://akama.arc.nasa.gov/ASRSDBOnline/QueryWizard_Filter.aspx.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[6] “all-mpnet-base-v2.” sentence-transformers/all-mpnet-base-v2, Hugging Face. Date Accessed: June 30, 2023. Online: https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[7] “all-distilroberta-v1.” sentence-transformers/all-distilroberta-v1, Hugging Face. Date Accessed: June 30, 2023. Online: https://huggingface.co/sentence-transformers/all-distilroberta-v1.

[8] “all-minilm-l6-v2.” sentence-transformers/all-MiniLM-L6-v2, Hugging Face. Date Accessed: June 30, 2023. Online: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

[9] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, “MPNet: Masked and Permuted Pre-training for Language Understanding,” *arXiv preprint arXiv:2004.09297*, 2020.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[11] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized autoregressive pretraining for language understanding,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[12] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using siamese BERT-networks,” *arXiv preprint arXiv:1908.10084*, 2019.

[13] “Pretrained Models.” Date Accessed: August 21, 2023. Online: https://www.sbert.net/docs/pretrained_models.html.

[14] 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, p. 143, 2020.

[15] A. Miyamoto, M. V. Bendarkar, and D. N. Mavris, “Natural language processing of aviation safety reports to identify inefficient operational patterns,” *Aerospace*, vol. 9, no. 8, p. 450, 2022.

[16] A. Chanen, "Deep Learning for Extracting Word-Level Meaning from Safety Report Narratives," in *2016 Integrated Communications Navigation and Surveillance (ICNS)*, IEEE, Apr. 2016.

[17] M. Seale, G. Nabholz, A. Ruvinsky, L. Walker, S. Abdullah, A. Strelzoff, D. Martinez, A. Hines, W. Bond, G. George, J. Church, O. Eslinger, D. Wade, A. Wilson, and N. Rigoni, "Unlocking Insights to Black Hawk Maintenance Data Using Innovative Big Data Analytics and Management Techniques," *ERS Journal*, vol. 1, no. 1, 2019.

[18] Y. Luo and H. Shi, "Using lda2vec topic modeling to identify latent topics in aviation safety reports," in *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, pp. 518–523, IEEE, 2019.

[19] X. Zhang, P. Srinivasan, and S. Mahadevan, "Sequential deep learning from ntsb reports for aviation safety prognosis," *Safety science*, vol. 142, p. 105390, 2021.

[20] T. Madeira, R. Melício, D. Valério, and L. Santos, "Machine learning and natural language processing for prediction of human factors in aviation incident reports," *Aerospace*, vol. 8, no. 2, p. 47, 2021.

[21] S. Kierszbaum, T. Klein, and L. Lapasset, "ASRS-CMFS vs. RoBERTa: Comparing Two Pre-Trained Language Models to Predict Anomalies in Aviation Occurrence Reports with a Low Volume of In-Domain Data Available," *Aerospace*, vol. 9, no. 10, p. 591, 2022.

[22] C. Chandra, X. Jing, M. V. Bendarkar, K. Sawant, L. Elias, M. Kirby, and D. N. Mavris, "Aviation-BERT: A Preliminary Aviation-Specific Natural Language Model," in *AIAA AVIATION 2023 Forum*, p. 3436, 2023.

[23] P. Jonk, V. de Vries, R. Wever, G. Sidiropoulos, and E. Kanoulas, "Natural language processing of aviation occurrence reports for safety management," in *Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022)*, (Dublin, Ireland), pp. 2015–2023, 2023.

[24] OpenAI, 2023. ChatGPT (July Version) [based on the GPT-3.5 architecture, Large Language Model]. https://chat.openai.com.