Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
©2023 ESREL2023 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P233-cd

Human Factor Identification in Aviation Accidents Using Contextual Word Embeddings

July Bias Macêdo

Center for Risk Analysis, Reliability Engineering and Environmental Modeling (CEERMA). Department, Department of Industrial Engineering, Federal University of Pernambuco (UFPE), Country: Brazil. E-mail: july.bias@ufpe.br

Plínio M. S. Ramos

CEERMA. Department, Department of Industrial Engineering, UFPE, Country: Brazil. E-mail: plinio.marcio@ufpe.br

Caio Souto Maior

CEERMA. Department, Department of Industrial Engineering, UFPE, Country: Brazil. E-mail: caio.maior@ufpe.bt

Márcio J. C. Moura

CEERMA. Department, Department of Industrial Engineering, UFPE, Country: Brazil. E-mail: marcio.cmoura@ufpe.br

Isis Didier Lins

CEERMA. Department, Department of Industrial Engineering, UFPE, Country: Brazil. E-mail: isis.lins@ufpe.br

Human error is a leading cause of aviation accidents and can result in significant loss of life. To support decision-making, the aviation industry collects extensive data, including written accident investigation reports that contain valuable information for risk analysis and accident management. Natural Language Processing (NLP) can assist experts in processing and analyzing these reports, enabling effective risk management and the proposal of preventive measures. This paper proposes a novel methodology for identifying human factors leading to aviation accidents using topic modeling based on contextual word-vector representations extracted from pre-trained Bidirectional Encoder Representation from Transformers (BERT). The proposed approach differs from previous studies identified in a systematic literature review. This methodology can provide useful insights for proposing preventive measures and training plans to reduce the risk of human error.

Keywords: Aviation accidents, human factors, topic modelling, aviation safety, natural language processing.

1. Introduction

Numerous industries and societies invest significant resources into the management of risks. Despite this, risk management poses various complex challenges and issues, particularly in relation to the foundation and implementation of Risk Analysis (RA). RA encompasses hazard identification, cause and consequence analysis, and risk assessment, enabling effective risk management and the prevention of potential accidents [1]. Consequently, any occurrence resulting from an unsafe act or condition can lead to adverse outcomes such as property damage, economic and social disruption, environmental degradation, and human fatalities. [2].

Accidents in the aviation industry have far-reaching consequences, causing both significant financial losses and loss of life [3]. The predominant cause of approximately 80% of such incidents is attributed to human error [4]. Indeed, despite the overall decrease in yearly accidents over the past few decades, Madeira et al., (2021) observed a shift in the primary latent cause of incidents towards human factors.

Furthermore, the aviation industry gathers data from diverse sources in varying formats, such as voice recordings from air traffic control, flight data from onboard measurement devices, and written accident investigation reports [6,7]. Accident investigation, a safety technique aimed at identifying and reporting the causes of a given accident, is critical in preventing future accidents and similar incidents through internal reporting and investigation. The significance of a robust investigation lies in the potential to derive preventive insights from past unexpected events [8,9].

In compliance with the regulations set forth by national and international regulatory bodies, companies are required to maintain a comprehensive collection of accident reports to enable safety professionals to analyze and address identified root causes [10,11]. However, the sheer volume of reports, which are typically written in natural language, makes it nearly impossible for human review of the entire database [12]. By leveraging Natural Language Processing (NLP) techniques, information from the text can be extracted, organized, and classified, allowing for the automatic identification of patterns[13]. As a result, the application of NLP techniques seems a promising solution to support RA.

The present study employed topic modelling utilizing contextual word embeddings extracted from pre-trained BERT [14] to identify human factors associated with aviation accidents. The proposed model was applied to a public database comprising accident investigation reports conducted by the National Transportation Safety Board (NTSB) between 1982 and 2022 [15].

2. Literature Review

The current body of literature (e.g., Ahmadpour-geshlagi et al., 2020; Andrzejczak et al., 2014; Baker et al., 2020; Hughes et al., 2018; Lombardi et al., 2019; Muguro et al., 2020; Single et al., 2020; Stephen and Labib, 2018) has utilized NLP techniques to extract information from accident investigation reports; however, certain limitations remain. For example, Kuhn (2019), applied Latent Dirichlet Allocation (LDA) topic modelling to identify known issues causing aviation accidents. Nevertheless, LDA represents a document as a Bag-of-Words (BoW) and does not account for the contextual relationships between words in a sentence. The author identified generic topics and failed to conduct a detailed analysis of the causes of accidents. Moreover, the author identified the human factor as the most represented topic but did not elaborate on the factors involved. In Yildiz et al. (2017), pilot fatigue was identified as the leading cause of aviation accidents, and the authors focused on modeling and analyzing fatigue. However, different types of human errors may lead to an accident, and identifying these factors offers a broad understanding of potential hazards, enabling the implementation of preventive measures to enhance safety.

By utilizing NLP techniques, text data can be analyzed computationally with minimal manual effort. In the realm of aviation safety, there is a paucity of research investigating and identifying the human factors contributing to accidents through NLP using accident narratives. To address this gap, we employ topic modeling techniques, utilizing contextual word-vector representations from pre-trained BERT [14], to identify human factors leading to aviation accidents.

3. Methodology

The proposed methodology can be summarized as follows: first, we performed data analysis, where the information contained in the aviation database is analyzed and the scope of the study is defined, selecting the relevant information. Next, we performed topic modelling, where we grouped similar accident causes, extracted the main topic of each group, and identified the human factor related to each accident group.

3.1. Aviation Accidents Data Analysis

The National Transportation Safety Board (NTSB) is responsible for investigating and reporting on accidental transportation events, including aviation accidents, certain road accidents, marine accidents, and the release of hazardous materials that occur during transportation. The NTSB aviation accident database contains factual information obtained from completed investigations of US civil aviation accidents and incidents. Initially, a preliminary report is made available online shortly after the accident, and this is later replaced by a final report detailing the accident and its probable cause. The NTSB database is accessible to the public, and our analysis involved examining aviation accident investigation reports conducted from 1982 to October 2022 [15].

The database encompasses over 20,000 documents that are organized into spreadsheets, providing extensive details about the accident, including the date of occurrence, the number of people who sustained injuries, the severity of injuries, and whether the individuals involved were passengers or crew members. Although the database covers a wide range of information, our primary focus is on the “accident description”, which provides a more detailed narrative of the incident, including information related to its cause and effects. The database also contains a separate, shorter, narrative referred to as the “accident cause description”, which outlines the cause of the accident. However, this information is not consistently available and lacks standardization, making it challenging for experts to extract insights about common causes of accidents and develop effective preventive measures.

Raw text data often includes irrelevant information that can affect the accuracy of predictive models, such as punctuation, stopwords, and typos. Therefore, text preprocessing is a crucial step in improving data quality by removing unwanted information, homogenizing documents, and reducing computational costs [5]. Text preprocessing involves a set of operations that are applied to textual data to eliminate noise and standardize the format. This is particularly important because text data often contain special characters, numbers, and dates that can add to the noise.

In this study, we performed text preprocessing on the “accident descriptions” and “accident cause descriptions” using both string methods and functions from the NLTK library [26]. The aim was to improve the data quality and reduce noise in the textual data before performing further analysis. Three preprocessing operations were applied: stop word filtering, lowercasing, and tokenization. Stop word filtering was used to remove non-informative terms such as “the”, “it”, and “is”. The terms were also converted to lowercase to standardize the text. Finally, tokenization was applied to split the text into individual words, or tokens, allowing for further analysis and processing.

Once the preprocessing step was completed, we filtered the database to select only those accidents that contained both the event description and the “accident cause description”. Furthermore, we narrowed our focus to accidents related to human error, as we aim to identify human factors. To automatically identify such accidents, we searched for the expression ‘pilot's failure’ in the descriptions, resulting in a dataset of 10,530 accidents. This approach ensured that the selected accidents were relevant to our research question and allowed us to proceed with further analysis of the data.

To prepare the preprocessed accident descriptions for fine-tuning with the pre-trained BERT model, we added special tokens [CLS] and [SEP] to mark the beginning and end of each description. The BERT model was originally pre-trained using such a format, making it necessary to use this format for fine-tuning. To achieve this, we utilized the ‘AutoTokenizer’ from the transformers library, which splits the sentences into a sequence of tokens based on punctuation and sub-word units, maps vocabulary tokens to indices, and converts raw text to sparse index encodings. The cleaned sentences were processed by the tokenizer, and sequences were padded with zeros to a maximum length since the BERT model requires inputs that have the same shape and size.

3.2. Topic Modelling

Accident descriptions often contain information about the cause of the incident, but extracting this information manually can be a time-consuming task for experts. To address this issue, an automatic classifier could be developed to identify accident causes from the descriptions, allowing experts to evaluate them more efficiently and uncover common patterns. However, training such a classifier requires a labelled database of accidents, which can be challenging to create. In this study, we utilized topic modeling on the “accident cause descriptions” to identify common causes and label the database more efficiently. It is worth noting that the NTSB dataset used in this research provides cause descriptions rather than categorized causes.

Topic modeling techniques can be a valuable resource for revealing common themes within a set of documents. Popular models, such as LDA and its variants [27], characterize a document as a BoW and model each document as a combination of underlying topics. However, BoW representations do not consider the context of words within a sentence. In contrast, BERT [14] has demonstrated impressive performance in generating context-based word-vector and sentence-vector representations [28]. In light of this, we employed a topic modeling strategy based on [29].

Our analysis pipeline first utilizes a pre-trained language model to convert each “accident cause description” into its corresponding embedding representation, capturing document-level information. We then reduce the dimensionality of these embeddings through Uniform Manifold Approximation and Projection (UMAP) [30], a technique which optimizes the clustering process. Using Hierarchical Density-Based Cluster Analysis (HDBSCAN) [31], we group semantically similar accident causes into clusters, each representing a distinct topic. Finally, we extract topic representations from the resulting clusters of documents using a variant of the widely used TF-IDF [32] technique.

Accidents cause descriptions are embedded in vector space that can be compared semantically. We assume that documents containing the same topic are semantically similar. To perform the first step, we loaded the Pytorch implementation of pre-trained BERT, specifically the ‘BertForSequenceClassification’ model, available at the transformers library [33]. We inputted the accident cause descriptions into the BERT model and extracted feature vectors for each word from the last layer. This step allowed us to represent each description as a dense vector in the BERT embedding space, capturing its semantic meaning.

To put it simply, each word in the accident cause description is represented as a vector with 768 features. As a result, each document, comprising n words, is represented by n 768-dimensional vectors. However, for the clustering process, each document must be represented by a single vector. To achieve this, we compute the mean of the n embedding vectors to obtain a vector representation for each document.

To reduce the dimensionality of the 768-dimensional vectors representing each word in the accident cause descriptions, we used UMAP. Although t-SNE [34] and PCA [35] are commonly used for dimensionality reduction, UMAP offers a better balance between preserving the global structure of the data and running time when projecting to lower dimensions [36].

In the next step we performed unsupervised clustering using HDBSCAN. HDBSCAN is advantageous because it can automatically identify the number of clusters, or partitions, without requiring this information as input. Additionally, HDBSCAN is capable of identifying dense clusters, and it does not require that every data point is assigned to a cluster. Outliers, or data points that do not belong to any cluster, can be detected as well [31,37]. The scikit-learn library was used to implement both UMAP and HDBSCAN in this study [38].

Next, a class-based variation of TF-IDF is applied to the set of clusters obtained. Specifically, TF-IDF formula (Eq. 2) is adopted for multiple classes by joining multiple documents per class.

$$W_{w,c} = |tf_{w,c}| \times \log\left(1 + \frac{A}{f_w}\right) \tag{1}$$

The given formula uses $tf_{w,c}$, which represents the frequency of each word w in each class c after normalization. The value of A represents the total number of documents, and f_w is the document frequency of word w across all classes. Thus, each cluster is converted to a single document instead of a set of documents.

By assigning a single importance value $W_{w,c}$ to each word w in a cluster c, we can generate a topic. This approach allows us to identify the most significant n words in each cluster, which provide a reliable representation of the topic. In this study, we selected the top 5 words to describe the primary topic of each cluster. Using this method, we evaluated the factors that contribute to accidents caused by human error and labeled the accidents based on the identified human factors.

5. Results

To prepare for cluster analysis, we transformed the preprocessed “accident cause descriptions” associated with pilot errors into embeddings. The embeddings were initially comprised of 768 dimensions, but we reduced them to just 5 using UMAP. Next, we utilized the HDBSCAN algorithm to fit the reduced embeddings, which resulted in 98 clusters. However, due to the large number of clusters, we merged them together by means of cosine similarity. This allowed for easier and more efficient cluster analysis [37].

Thus, theanalysis resulted in a total of 11 clusters, with one of them being labeled as '-1'. This cluster contained 5,913 instances which were not assigned to any specific cluster. Fig. 1 illustrates the distribution of instances across clusters 0 to 9, and it can be observed that the clusters are fairly balanced with a combined total of 4,617 instances. To uncover the main topic of each cluster, we applied TF-IDF to the data and extracted the top 5 most significant words. In the upcoming section, we will delve into the topics identified in each cluster and correlate them with the respective human factors.

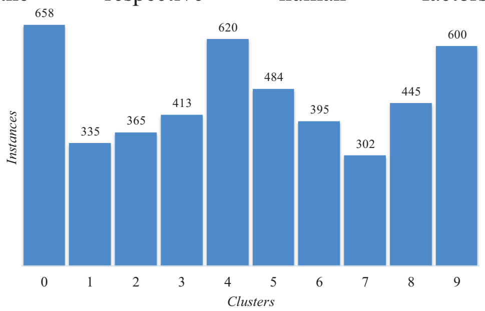

Fig. 1. The number of instances per cluster.

In the next section, we will present the analysis of the topics in each cluster and the identification of the corresponding human factors.

5.1. *Topic Analysis*

In Table 1, we present the top 5 words associated with each cluster. It is evident from the topics that the '-1' cluster does not depict a specific accident cause or human factor. This is because the words that represent this cluster are general and likely to be present in all "accident cause descriptions". For instance, we can expect words such as 'airplane' and 'flight' to appear in all accidents, or that the described events ultimately 'resulted' in an accident or consequence. Our next step was to examine the topics identified in clusters 0 to 9 and determine the human factors responsible for each accident. Our aim is to construct a labeled dataset that can be used to train a classifier to recognize descriptions of accidents associated with human factors.

Table 1. The top 5 words of each cluster were obtained using TF-IDF.

Cluster	Top 5 words				
	1	2	3	4	5
-1	Airplane	Flight	Resulted	Contribution	Runway
0	Directional	Roll	Control	Maintain	Take off
1	Flare	Hard	Improper	Landing	Student
2	Clearance	Manoeuvring	Lines	Low	Maintain
3	Bounced	Flare	Improper	Recovery	Hard
4	Directional	Excursion	Runway	Loop	Roll
5	Aerodynamic	Stall	Airspeed	Adequate	Attack
6	Gusting	Conditions	Cross wind	Wind	Directional
7	Fuel	Exhaustion	Engine	Power	Total
8	Compensation	Inadequate	Contributing	Accident	Cross wind

Our clustering analysis revealed that clusters 0 and 4 are characterized by accidents caused by loss of directional control. The term 'directional' was found to be the most significant word in both clusters, strongly indicating this type of accident. Moreover, the words 'control' and 'maintain' in cluster 0 were found to be directly related to loss of directional control, suggesting a clear relationship between them. In cluster 4, the terms 'runway' and 'excursion' were also associated with this type of accident. Based on these findings, we labeled the accidents in clusters 0 and 4 with the human factor 'loss of directional control'.

Cluster 1 accidents were labeled with 'maneuver/action failure'. The main topic of cluster 1 appears to be related to improper actions or maneuvers. The occurrence of words like 'flare', 'hard', and 'landing' along with 'improper' may indicate improper actions that result in a hard landing and/or improper flare/landing. The top words in cluster 2 suggest that accidents occur due to obstacles, as evidenced by the words 'lines' and 'clearance', which refer to power lines and maintaining clearance from objects. Additionally, the word 'low' is likely related to the aircraft's altitude, which combined with the word ‘maintain’, is a good indication that the accidents are related to a failure to maintain the correct altitude to avoid the obstacle. As a result, accidents in cluster 2 were labeled with ‘obstacle clearance failure’.

The top words in cluster 5, such as ‘stall’, ‘adequate’, and ‘airspeed’, are highly specific and suggest failure to maintain appropriate stall speed. Therefore, the accidents in cluster 5 were tagged with the human factor ‘airspeed control failure’. It appears that clusters 6 and 8 share similar themes, so we opted to assign the same human factor to accidents in these clusters (similar to what was done with clusters 0 and 4). The top words in these clusters indicate that the accidents were caused by gusty winds and crosswind conditions. Additionally, words like ‘compensation’, ‘inadequate’, and ‘adequate’ lead us to conclude that the accidents were due to failure to compensate for such conditions.

Cluster 7 is mainly associated with accidents that are linked to the fuel level. The key term in this cluster is ‘fuel’, and other terms such as ‘exhaustion’, ‘loss’, ‘power’, and ‘engine’ further indicate situations where low fuel levels may lead to engine power loss, for instance, when the pilot overlooks or fails to maintain the appropriate fuel level. As a result, the accidents in cluster 7 were classified as ‘failure to maintain fuel level’.

Finally, cluster 9’s topic does not seem to be related to human error. Its primary words include ‘carburetor’ and ‘engine’, indicating that accidents may be due to equipment malfunction. As the purpose of the suggested approach is to recognize human factors, clusters -1 and 9 will be disregarded. This determination is reasonable because the methodology aims to label the accident database in a clear, direct, and mostly automated way, meaning that it is unnecessary to evaluate each case’s descriptions individually. Thus, the analysis of the topics resulted in the identification of seven ‘Human Factor’, which are summarized in Table 2.

Table 2. Human factors identified after topic analysis.

Cluster	Human Factor
0 and 4	Control failure
1	Action failure
2	Obstacle clearance failure
3	Landing recovery failure
5	Airspeed control failure
6 and 8	Flight conditions compensation failure
7	Failure to maintain fuel

6. Conclusion

In this paper, we have presented a methodology that facilitates the identification of human factors associated with aviation accidents. Since human factors are a major cause of such accidents, accident investigation reports provide valuable information for making informed decisions and identifying causes to prevent recurrence. Our proposed methodology involves labeling the accident dataset, which is crucial for developing the classifier, with minimal effort and without manual analysis of each “accident description”. We utilized contextual embeddings and topic modeling to identify human factor categories and labeled the accidents accordingly. We used a public database called NTSB, which allowed us to identify and summarize the categories into seven separate clusters. Our contributions are expected to assist experts in identifying common causes of human error, providing insights to devise preventive measures and training programs to mitigate the risk of human failure.

The labeled dataset obtained could be used to train a classifier. This could prove to be helpful as accident descriptions are usually lengthy, and automatically identifying the human factors responsible for the accident could save time for experts during the analysis process. Additionally, the proposed methodology is straightforward to implement and can be applied periodically to identify any new human factors and retrain the classifier. This ensures that the classifier remains current and up-to-date at all times.

Acknowledgements

The authors thank CNPq, FACEPE, and PRH 38.1 managed by ANP and FINEP for the financial support through research grants. This study was financed in part by CAPES – Finance Code 001.

References

[1] ISO, ISO 31000: Risk management — Guidelines, (2018).

[2] P.M. Ramos, J.B. Macedo, C.B. Maior, M.C. Moura, I.D. Lins, Combining BERT with numerical variables to classify injury leave based on accident description, Proc. Inst. Mech. Eng. Part O J. Risk Reliab. (2022) 1748006X2211401. https://doi.org/10.1177/1748006X221140194 .

[3] M. Studic, A. Majumdar, W. Schuster, W.Y. Ochieng, A systemic modelling of ground

handling services using the functional resonance analysis method, Transp. Res. Part C Emerg. Technol. 74 (2017) 245–260. https://doi.org/10.1016/j.trc.2016.11.004.

[4] D. Abdullah, H. Takahashi, U. Lakhani, Improving the Understanding between Control Tower Operator and Pilot Using Semantic Techniques - A New Approach, in: Proc. - 2017 IEEE 13th Int. Symp. Auton. Decentralized Syst. ISADS 2017, IEEE, 2017: pp. 207–211. https://doi.org/10.1109/ISADS.2017.8.

[5] T. Madeira, R. Melício, D. Valério, L. Santos, Machine Learning and Natural Language Processing for Prediction of Human Factors in Aviation Incident Reports, Aerospace. 8 (2021) 47. https://doi.org/10.3390/aerospace8020047.

[6] A. Miyamoto, M. V. Bendarkar, D.N. Mavris, Natural Language Processing of Aviation Safety Reports to Identify Inefficient Operational Patterns, Aerospace. 9 (2022). https://doi.org/10.3390/aerospace9080450.

[7] F. Gürbüz, L. Özbakir, H. Yapici, Data mining and preprocessing application on component reports of an airline company in Turkey, Expert Syst. Appl. 38 (2011) 6618–6626. https://doi.org/10.1016/j.eswa.2010.11.076.

[8] S. Jones, C. Kirchsteiger, W. Bjerke, The importance of near miss reporting to further improve safety performance, J. Loss Prev. Process Ind. 12 (1999) 59–67. https://doi.org/10.1016/S0950-4230(98)00038-2.

[9] F. Salguero-Caparros, M. Suarez-Cebador, J.C. Rubio-Romero, Analysis of investigation reports on occupational accidents, Saf. Sci. 72 (2015) 329–336. https://doi.org/10.1016/j.ssci.2014.10.005.

[10] F. Abdat, S. Leclercq, X. Cuny, C. Tissot, Extracting recurrent scenarios from narrative texts using a Bayesian network: Application to serious occupational accidents with movement disturbance, Accid. Anal. Prev. 70 (2014) 155–166. https://doi.org/10.1016/j.aap.2014.04.004.

[11] R. Bavaresco, H. Arruda, E. Rocha, J. Barbosa, G.-P. Li, Internet of Things and occupational well-being in industry 4.0: A systematic mapping study and taxonomy, Comput. Ind. Eng. 161 (2021) 107670. https://doi.org/10.1016/j.cie.2021.107670.

[12] S.J. Bertke, A.R. Meyers, S.J. Wurzelbacher, J. Bell, M.L. Lampl, D. Robins, Development and evaluation of a Naïve Bayesian model for coding causation of workers' compensation claims, J. Safety Res. 43 (2012) 327–332. https://doi.org/10.1016/j.jsr.2012.10.012.

[13] B. Drury, M. Roche, A survey of the applications of text mining for agriculture, Comput. Electron. Agric. 163 (2019) 104864. https://doi.org/10.1016/j.compag.2019.104864.

[14] J. Devlin, M. Chang, L. Kenton, T. Kristina, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ArXiv Prepr. ArXiv1810.04805. (2018).

[15] NTSB, Accident Data, Download Census US Civ. Aviat. Accid. (2022). https://data.ntsb.gov/avdata (accessed August 1, 2022).

[16] R. Ahmadpour-geshlagi, N. Gillani, S. Azami–Aghdash, M. Javanmardi, S.S. Alizadeh, S. Jalilpour, Investigating the status of accident precursor management in East Azarbaijan Province Gas Company, Int. J. Occup. Saf. Ergon. (2020) 1–12. https://doi.org/10.1080/10803548.2020.1770451.

[17] C. Andrzejczak, W. Karwowski, W. Thompson, The Identification of Factors Contributing to Self-Reported Anomalies in Civil Aviation, Int. J. Occup. Saf. Ergon. 20 (2014) 3–18. https://doi.org/10.1080/10803548.2014.11077029.

[18] H. Baker, M.R. Hallowell, A.J.P. Tixier, Automatically learning construction injury precursors from text, Autom. Constr. 118 (2020) 103145. https://doi.org/10.1016/j.autcon.2020.103145.

[19] P. Hughes, D. Shipp, M. Figueres-Esteban, C. van Gulijk, From free-text to structured safety management: Introduction of a semi-automated classification method of railway hazard reports to elements on a bow-tie diagram, Saf. Sci. 110 (2018) 11–19. https://doi.org/10.1016/j.ssci.2018.03.011.

[20] M. Lombardi, M. Fargnoli, G. Parise, Risk profiling from the European statistics on accidents at work (ESAW) accidents′ databases: A case study in construction sites, Int. J. Environ. Res. Public Health. 16 (2019). https://doi.org/10.3390/ijerph16234748.

[21] J.K. Muguro, M. Sasaki, K. Matsushita, W. Njeri, Trend analysis and fatality causes in Kenyan roads: A review of road traffic accident data between 2015 and 2020, Cogent Eng. 7 (2020). https://doi.org/10.1080/23311916.2020.1797981.

[22] J.I. Single, J. Schmidt, J. Denecke, Knowledge acquisition from chemical accident databases using an ontology-based

method and natural language processing, Saf. Sci. 129 (2020) 104747. https://doi.org/10.1016/j.ssci.2020.104747.

[23] C. Stephen, A. Labib, A hybrid model for learning from failures, Expert Syst. Appl. 93 (2018) 212–222. https://doi.org/10.1016/j.eswa.2017.10.031.

[24] K.D. Kuhn, Using structural topic modeling to identify latent topics and trends in aviation incident reports, Transp. Res. Part C. 87 (2019) 105–122. https://doi.org/10.1016/j.trc.2017.12.018.

[25] B.C. Yildiz, F. Gzara, S. Elhedhli, Airline crew pairing with fatigue: Modeling and analysis, Transp. Res. Part C Emerg. Technol. 74 (2017) 99–112. https://doi.org/10.1016/j.trc.2016.11.002.

[26] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, O'Reilly Media, Inc., 2009.

[27] R. Srivastava, P. Singh, K.P.S. Rana, V. Kumar, A topic modeled unsupervised approach to single document extractive text summarization, Knowledge-Based Syst. 246 (2022) 108636. https://doi.org/10.1016/j.knosys.2022.108636 .

[28] Q. Liu, M.J. Kusner, P. Blunsom, A Survey on Contextual Embeddings, ArXiv. arXiv prep (2020). http://arxiv.org/abs/2003.07278.

[29] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, ArXiv Prepr. arXiv:2203 (2022). http://arxiv.org/abs/2203.05794.

[30] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv Prepr. arXiv:1802 (2018). http://arxiv.org/abs/1802.03426.

[31] I. Ghamarian, E.A. Marquis, Hierarchical density-based cluster analysis framework for atom probe tomography data, Ultramicroscopy. 200 (2019) 28–38. https://doi.org/10.1016/j.ultramic.2019.01.011.

[32] L. Havrlant, V. Kreinovich, A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation), Int. J. Gen. Syst. 46 (2017) 27–36. https://doi.org/10.1080/03081079.2017.1291635.

[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, M. Funtowicz, J. Davison, S. Shleifer, P. Von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A.M. Rush, Transformers : State-of-the-Art Natural Language Processing, in: Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. Syst. Demonstr., Association for Computational Linguistics, 2020: pp. 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[34] W. Li, J.E. Cerise, Y. Yang, H. Han, Application of t-SNE to human genetic data, J. Bioinform. Comput. Biol. 15 (2017) 1–14. https://doi.org/10.1142/S0219720017500172.

[35] S. Marukatat, Tutorial on PCA and approximate PCA and approximate kernel PCA, Springer Netherlands, 2022. https://doi.org/10.1007/s10462-022-10297-z.

[36] M. Bahri, B. Pfahringer, A. Bifet, S. Maniu, Efficient Batch-Incremental Classification Using UMAP for Evolving Data Streams, in: Adv. Intell. Data Anal. XVIII, 2020: pp. 40–53. https://doi.org/10.1007/978-3-030-44584-3_4.

[37] G. Stewart, M. Al-Khassaweneh, An Implementation of the HDBSCAN* Clustering Algorithm, Appl. Sci. 12 (2022) 1–21. https://doi.org/10.3390/app12052405.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.