International Conference on Military Technologies (ICMT) 2019
May 30 – 31, 2019, Brno, Czech Republic

Automated Aviation Occurrences Categorization

Kosio Marev and Krasin Georgiev†
†Department of Aeronautics, Technical University of Sofia, 1000 Sofia, Bulgaria, e-mail: *krasin@tu-sofia.bg*

***Abstract*— Information about aviation events is collected by all participants in the aviation system, e.g. airlines, maintenance organizations, and air traffic controllers. Reporting and initial assessment usually involves assigning categories from a predefined nomenclature (scheme) aligned with the purpose of the reporting system and the established processing practices. Such manual categorization is time and resource consuming and, more importantly, limiting the application of the dataset. We apply and evaluate the effectiveness of a state of the art Neural Networks based algorithm for Natural Language Processing for classification of aviation safety report narratives. Multi-class, multi-label supervised learning is performed on two small datasets, 4500 and 8000 cases with 16 and 54 classes respectively, both extracted from the NASA Aviation Safety Reporting System. The results are promising if compared to recent studies and considering that an off the shelf algorithm without much customization is applied. Automatic categorizations can relief the current burden for manual categorization of the events by reducing the number of likely categories, targeting quality checks to most ambiguous records and applying new or updated classification schemes.**

***Keywords*— aviation safety, occurrence reporting, NLP, NASA ASRS, fastai, neural networks**

I. Introduction

Multiple aviation safety information collection programs are in operation today [1]. They accumulate shared experience of flight, operational, maintenance, and regulatory staff either as individuals or organizations. Some of the reports are mandatory, required from the regulations to monitor the risk from known safety threats and other are voluntary, allowing unrestricted exchange of personal observations and safety concerns [2]–[5]. In most cases, measured numerical parameter, event characteristics and categories (assigned originally or after additional assessments) are supplemented by a free form event description provided as simple text.

What information is collected and how it is coded into categories is mostly determined by the purpose of each database and the established analysis practices. However the reporting and analysis requirements can change in time. For example, the European Aviation Safety Agency (EASA) has begun applying a new European Risk Classification Scheme (ERCS) to historical occurrences [3], [6]; the Human Factor Analysis and Classification System (HFACS) has been applied to study past accidents [7], [8]. Applications other than the initial design are severely restricted by the manpower required to do the necessary knowledge extraction. Careful reading, understanding, assessment and final labeling or number extraction of hundreds of thousands of narratives by experts is rarely an option.

Automated natural language processing (NLP) of incident and accident reports for aviation risk management has been studied in the recent years [9]–[14]. Some success has been demonstrated on tasks as multi-class and multi-label categorization, topic modeling and similarity search. Interactive browsing and exploration approach was proposed in [12]. Practical applications of the automatic classification discussed are

- reducing the number of possible categories in new events categorization for current databases;
- using customized nomenclatures to re-categorize old datasets for specific tasks as reliability and risk analysis;
- an interactive search based on keywords or sample reports.

A short overview of the techniques adopted for classification of aviation narratives follows. The basic approach splits the problem in two tasks, first represent the text by an array of numbers with fixed length (feature extraction) and then transform the features into one or more labels (classification).

Each document is converted to a vector of text unit frequencies ("bag" of words, e.g. see [11]). The document vector is with fixed length equal to the number of text units in the vocabulary of the dataset. Each value in the vector is the frequency of the corresponding term in the document. Hand-written, rule-based normalizers has been used to convert synonyms, abbreviations, and multi-word terms into single terms, numbers to signifiers, etc. in [11]. Then different types of text units (also called tokens or terms) are constructed, from single terms or stems to n-grams of terms or characters. Each term frequency (TF) can be scaled to account the "rarity" of the term in the collection, the so called TF*IDF representation, where IDF is for Inverse Document Frequency. Some approaches further apply Latent Semantic Analysis (LSA) to reduce the document vector from vocabulary word frequencies (several thousands of terms) to topic frequencies (several hundred semantic topics) [13]–[15]. The Document Terms matrix is decomposed into Document-Topic and Topic-Terms matrices using appropriate matrix factorization. This is an unsupervised learning method for feature dimension reduction and topic extraction. Probabilistic alternatives for topic modelling exist [16], [17], and were applied for aviation safety reports in [9].

Above studies does not exploit the modern representation of the documents as a sequence of word embeddings [18]–[20]. Each word is represented by a vector instead of an index in the vocabulary and similar words are closer in the vector space – the so called distributed representation.

978-1-7281-4593-8/19/$31.00

The classification part is done using conventional algorithms or newly developed ones. A correlation between document-term and topic-term vectors combined with threshold was applied in [10]; K-Nearest-Neighbor (KNN) classifier based on document-topic vectors cosine similarity in [13], [14]; Support Vector Machines (SVM) over document term vectors – [12]; Bayesian multi-variate regression on document-topic vectors – [9], etc. The multi-label text classification is handled in different ways [21], i.e. training independent classifiers for each target category ([12] – 37 SVM classifiers with linear kernels), splitting the document to sentences and taking the most frequent sentence classes (with KNN classifier in [14]), using a distribution classifier that can output distribution of probabilities for all labels [9].

Again it seems reasonable to consider the recent achievements in the field of machine learning and NLP. Many traditional algorithms in computer vision, machine translation and automatic control have been replaced by deep neural networks [22], [23]. The power of the new "data"-based approach is demonstrated by the advancements in self-driving cars, robotics and game industry. The need for big training dataset has been relaxed in image processing by the so called "transfer learning" – existing models trained for different problem and on unrelated data are fine-tuned for the particular task. Current researches develop transfer learning techniques for NLP by using pretrained word vectors to build document representation for further fine tuning and classification [24], [25]. Howard and Ruder propose pretraining and fine tuning of a whole language model and show that their method outperforms existing state-of-the-art on multiple representative text classification tasks [25].

The power and limitations of the neural networks (NN) based approach has not been demonstrated on aviation occurrence narrative data. The goal of this study is to apply an off the shelf NN NLP technique with minor modifications on reports from the NASA Aviation Safety Reporting System (ASRS) database. The algorithm is introduced in the next section and follows [25]. The achieved performance will be discussed in the context of similar metrics reported in the aviation safety literature [9], [12], [14].

II. Materials and Methods

Data for the training and validation datasets were taken from ASRS Database Online [26]. The query reproduced the filter applied by Robinson (2018) to allow fair comparison of the results [14]. The dataset included passenger flights under FAR Part 121, all records for years 2011 and 2012 for the training and for year 2009 for the validation subsets. January 2013 was also included in the training subset. This resulted in 4500 training and 2948 validation reports. Each report can have single primary problem and multiple contributing factors / situations. There are 16 possible labels. The primary and the contributing factors are treated as separate datasets. Above training dataset was expanded with additional 6242 records from years 2010, 2013 and 2014 to form an extra-training dataset with 10742 reports.

Another study based on ASRS narratives tries to predict the type of the anomalous events [9]. There are 57 predefined classes and a randomly selected subset of 10000 reports was explored. For our study, 10000 records from the expanded extra-training dataset described above were used to allow better reproducibility of the results. Training and validation datasets were produced using 5-fold cross-validation.

A variety of metrics can be used for performance evaluation of the model predictions [27]. We consider only the ones applicable for both multi-class and multi-label problems. Most metrics were calculated using the implementations in [27]. The performance metrics selected for this study follow the ones used in [14] and [9] to allow comparison:

- Hamming Loss (H_L) – the fraction of the wrong labels to the total number of labels. This is the least restrictive measure, as full credit is given for all matching records and labels combinations.

- Exact Match Ratio (A) – the fraction of records with exact match of all labels. This is the most restrictive measure, as no credit is given for correctly predicting some of the labels of a record.

- Precision (P) and Recall (R) scores – the fraction of true predicted positive to all predicted positive (precision) and to all positive (recall) respectively. These are appropriate for measuring the performance for unbalanced data. Several schemes of averaging exist to apply these metrics for multi-class and multi-label data – micro, macro, weighted and sample, i.e. averaging as a whole, averaging by labels, by labels with weighting based on label frequency, and by records, respectively.

- F1 scores – a harmonic mean of precision and recall, appropriate for measuring the performance for unbalanced data.

- Multilabel ranking metrics – label ranking average precision, coverage error and label ranking loss as defined and implemented in [9], [27]. These are metrics that do not require explicit prediction of the labels.

- One error – evaluates "how frequently the top ranked predicted label is not among the true labels" [9].

A method called Universal Language Model Fine-tunning (ULMFiT) was used for the classification [25]. The model is a sequence of a language model (LM) with 3-layer AWD LSTM architecture and a pooling linear classifier [25], [28]. The training procedure has three steps: 1) general LM pretraining; 2) fine tuning the LM; 3) classifier fine-tunning. All calculations were based on the code accompanying [25]. For the first step a pretrained model was imported. The fine tuning of the LM was done with the safety narratives from the dataset. The classifier in the third step was modified to allow training with multi-label input data. The cross entropy loss function was replaced with binary cross entropy with logits:

$$\mathcal{L} = \sum_{l=1}^{L} \left[x_l - x_l y_l + \log(1 + e^{-x_l}) \right], \qquad (1)$$

where is the output x_l for label l at the last linear layer without non-linear activation and y_l is the true value for label l, i.e. 1 or 0. The prediction matrix was calculated from the output scores x_l by threshold crossing. The threshold was selected based on the required precision-recall balance. As a whole, the model structure and model hyperparameters were preserved as in the original implementation [29].

The calculations were performed using Python (v.3.6.4, Anaconda Inc.) with Pandas (v.0.22.0), Numpy (v.1.14.2), SciKit-learn (v.0.19.1), spacy (v.2.0.11, Explosion AI), torch (v.0.4.1.post2) and Fastai (v. 0.7.0) libraries.

III. Results

Free text classification using a state-of-the-art off the shelf neural networks based approach was applied on safety report narratives datasets. The approach applies the so called Universal Language Model Fine-tuning (ULMFiT) method with transfer learning for text classification [25]. A language model based on recurrent neural network and a classifier based on fully connected network are available as open source pretrained model and code [25], [29]. The pretrained with Wikipedia texts language model was fine-tuned with safety records narratives. The safety report datasets were samples of narratives and the corresponding labels, e.g. primary problem / contributing factors, abnormal event types, etc. from the NASA ASRS database.

A. Data characteristics

The dataset of ASRS narratives with contributing factors will be explored in details. The frequency distribution of the labels of the documents is shown for both the training and the validation subsets in Table I. There are 4500 records in the train dataset and 2948 in the validation dataset. The number of unlabeled records is only 15 for the training and 13 for the validation parts so these can be safely ignored. The labels are selected from a set of 16 values without restriction of the number of labels per record. It is important to note that there is a data imbalance as some of the categories are rare with less than 3% of the observations while other are assigned to more than 50% of all cases ("Aircraft" and "Human factor"). This means that our classifier predictions for the former records will be poor but the latter will dominate most of the performance measures.

The multi-label problem can be characterized by the number of labels per record. It varies from zero (0.35 %) to ten (< 0.1 %) and most of the records have single label (36 %), two labels (30 %), three labels (20 %) and four labels (8 %). The average number of labels per record is 2.26. Predicting the exact number and combination of labels for each record is unnecessarily ambitious task and most studies rely on average precision, recall, and f1 scores as in [12], [14] or different ranking metrics [9]. To reduce the penalty from miss-represented labels, micro averaging was used. Other performance metrics were calculated also to allow discussions about individual categories and labels, and comparison with other studies.

TABLE I
DOCUMENT FREQUENCY BY LABEL. DATA FROM ASRS

	Train		**Val**	
Category	count	%	count	%
Aircraft	2841	63.1	1635	55.5
Human Factors	2264	50.3	1672	56.7
Procedure	1663	37.0	754	25.6
Company Policy	711	15.8	624	21.2
Weather	447	9.9	270	9.2
Environment [1]	446	9.9	238	8.1
Chart Or Publicat.	478	10.6	259	8.8
Airport	278	6.2	212	7.2
ATC [2]	107	2.4	60	2.0
Manuals	321	7.1	163	5.5
MEL	129	2.9	84	2.8
Equipment / Tooling	74	1.6	56	1.9
Part [3]	113	2.5	71	2.4
Airspace Structure	107	2.4	92	3.1
Staffing	71	1.6	52	1.8
Logbook Entry	101	2.2	59	2.0
Sum	23557	225.6	6301	213.7
Total records	4500	100	2948	100

[1] Environment – Non Weather Related;
[2] DTC Equipment / Nav Facility / Buildings;
[3] Incorrect / Not Installed / Unavailable Part

B. Classification performance

The results reported in Robinson 2018 for multi-label task [14] are replicated in Table II together with our performance metrics on the same datasets. On the multi-class and multi-label contributing factors dataset we reduce the Hamming loss by 54 % (from 0.216 to 0.099) and improve the F1 score by 37 % (from 0.484 to 0.663). In addition, two dummy predictions are shown to set a baseline for comparison – predicting all possible labels for each record (column 3) and always predicting the most frequent label "Aircraft" from the training dataset (Column 4).

TABLE II
PERFORMANCE METRICS FOR CONTRIBUTING FACTORS LABELS

Metric	[14] [1]	[14] [2]	dummy ones [3]	dummy fixed [4]	ULMFiT
Hamming loss	0.216	0.364	0.866	0.125	0.099
Precision	0.351	0.255	0.134	0.570	0.608
Recall	0.781	0.935	1.0	0.265	0.729
F1 score	0.484	0.400	0.237	0.362	0.663

[1] contributing by primary algorithm; [2] contributing by contributing algorithm [3] all labels as ones; [4] all zeros except "aircraft";

The results reported in Agovic 2010 for anomalous events labeling task [9] are replicated in Table III together with our performance metrics on a similar dataset (same database, same size of the dataset, 54 instead of 57

TABLE III

PERFORMANCE METRICS FOR EVENT TYPE LABELS

Metric	BMR [9]	ULMFiT
OneError	38.5 ± 0.8	27.0 ± 2.7
AvePrec	64.0 ± 0.5	75.2 ± 2.4
Coverage	8.17 ± 0.14	5.01 ± 0.75
HammingLoss	4.4 ± 0.0	2.9 ± 0.4
RankLoss	5.7 ± 0.2	3.2 ± 0.4

classes) and the same type of cross-validation. Again, the label ranking average precision is higher and the errors are lower for our approach.

Better predictions were reported in the literature for larger datasets. Tulechki applied classical NLP techniques supplemented by hand written rules for categorization of 136861 aviation accident and incident reports into 37 event type [12]. The micro-average precision, recall and F1-score were P=86.79%, R=74.08%, F1=79.15% respectively. These values can't be fairly compared to ours as the dataset is more than 20 times larger. Unfortunately ECCAIRS databases are not public. Moreover, we are interested in a limited number of training samples, from a few hundred to a few thousands, a number that can be easily prepared by a small team of experts in a reasonable time.

The prediction success was further evaluated at a label level for the contributing factors dataset (Table IV). The most frequent labels "Human Factors" and "Aircraft" have f1 score of about 0.8, but the rare ones as "Logbook Entry" and "Staffing" are not detected at all. The general properties of the model described in [12], as poor performance for rare classes and inherent difficulty with some classes, were observed for our model-data setup also. The f1 score for label "Weather" (0.63) is much better than the scores for "Procedure", "Chart or Publication" and "Environment – Non Weather Related" (0.53, 0.51 and 0.17 respectively) even though the support samples are similar in number (Table IV).

Further training on the expanded training dataset leads to some increase of the micro scores (f1 = 0.68). The increase is more prominent for the rear cases as shown in Figure 1 where the predictions of the updated model are evaluated on the same validation set. The macro averaged f1 score was improved from 0.36 to 0.40.

Automatic classification can be applied for preliminary screening of records for further manual assessment. Then the recall is important as it determines what part of target records will be missed. A tradeoff between precision and recall is easy to achieve in order to satisfy such user requirements. In Table II, our approach, while giving better overall score, has a lower recall compared to [14]. This is fixed by reducing the threshold for positive labels so the precision/recall is changed from 0.61/0.73 to 0.56/0.78 or 0.42/0.90. A recall score of 0.9 means that the classifier will catch 90% of the target records. Then a manual review will be needed to discard 58% of the filtered records.

TABLE IV

PERFORMANCE METRICS PER LABEL. DATA FROM ASRS DATABASE. CONTRIBUTING FACTORS. ULMFIT ALGORITHM

	f1-score	**precision**	**recall**	**support**
Human Factors	79.5	71.3	89.7	1672
Aircraft	84.7	77.1	94.1	1635
Procedure	52.5	37.2	89.1	754
Company Policy	57.7	65.6	51.4	624
Weather	62.5	64.0	61.1	270
Chart Or Publication	51.1	46.9	56.0	259
Environment	17.4	33.7	11.8	238
Airport	46.9	49.7	44.3	212
Manuals	47.1	42.6	52.8	163
Airspace Structure	13.2	50.0	7.6	92
MEL	40.0	43.7	36.9	84
Incorrect ... Part	16.5	30.8	11.3	71
ATC	3.0	14.3	1.7	60
Logbook Entry	0.0	0.0	0.0	59
Equipment / Tooling	3.5	100.0	1.8	56
Staffing	0.0	0.0	0.0	52
micro avg	**66.3**	**60.8**	**72.9**	**6301**
macro avg	36.0	45.4	38.1	6301
samples avg	69.1	68.0	78.9	6301
weighted avg	64.3	61.3	72.9	6301

Automatic categorization can be used as a backup or quality control procedure for important manual assessments. A manual review of improperly labeled records has revealed obvious inconsistencies in the original report coding [12]. Moreover perfect qualitative assessment is not expected even from human raters. A study specially designed to test human rater reliability in HFACS categorization gives Krippendorff Alpha values of 0.79 across the four tires and 0.67 across the 19 categories [30]. Having a tool to select cases for expert review is especially sensible when the number of records is in hundreds (1020 accidents analyzed in [8]) or even in thousands (e.g. 14086 general aviation accidents and incidents with 31491 aircrew casual factors in [7]).

IV. CONCLUSION

Current state-of-the-art neural network models as ULMFiT can be applied for classification of aviation safety narratives. We show that a model based on ULMFiT outperforms alternative models for classification of safety occurrence narratives for small training datasets. The f1 score 0.484 reported by Robinson [14] was increased to 0.663 (using exactly the same ASRS training and validation datasets). This result was achieved without parameter tuning and therefore further improvement can be expected.

The maturity and the accessibility of the tools for automatic text classification mean that such techniques should become a regular element in the toolbox of the aviation safety analyst. The neural networks approach is flexible enough to handle natively both multi-class and multi-label problems. Further improvements are expected

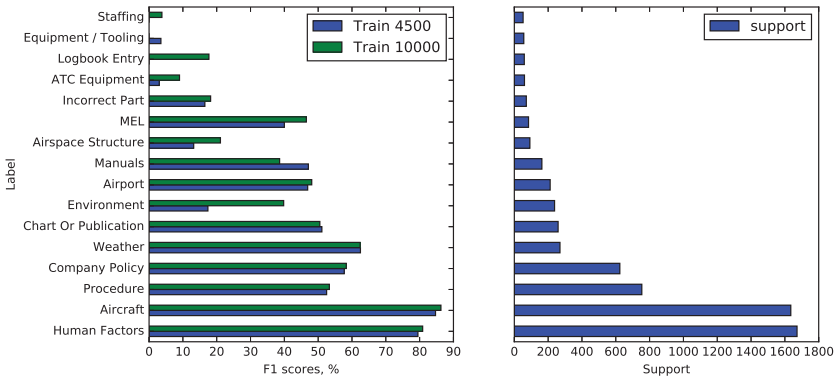

Fig. 1. Scores F1 by contributing factor label for training datasets with 4500 and 10000 samples. Support based on 2948 validation samples

as the neural network based NLP with transfer learning is an active field of research.

A next step is incorporating structured information to supplement the text processing algorithm. Pretraining the LM on large corpuses of aircraft technical documentation or other aviation related literature is another simple step to improve the model. Other NLP tasks as topic modeling (taxonomy evaluation) and similarity calculation (similarity based retrieval) could be studied.

REFERENCES

[1] GST, "Major Current or Planned Government Aviation Safety Information Collection Programs," p. 60, 2004.

[2] W. Reynard, C. E. Billing, E. S. Cheaney, and R. Hardy, "The Development of the NASA Aviation Safety Reporting System, NASA ASRS (Pub. 34)," NASA, Tech. Rep., 1986.

[3] EC, "Regulation (EU) No 376/2014 of the European Parliament and of the Council of 3 April 2014 on the reporting, analysis and follow-up of occurrences in civil aviation," pp. 18–43, 2014.

[4] CAA.UK, "CAP382: Occurrence Reporting Scheme," 2016. [Online]. Available: https://www.caa.co.uk/Our-work/Make-a-report-or-complaint/MOR/Occurrence-reporting/

[5] FAA, "SDR Reporting." [Online]. Available: https://av-info.faa.gov/sdrx/Default.aspx

[6] EASA, "EASA Annual Safety Review 2018," European Aviation Safety Agency, Safety Intelligence and Performance department, Cologne, Germany, Tech. Rep., 2018.

[7] S. A. Shappell and D. A. Wiegman, "A Human Error Analysis of General Aviation Controlled Flight Into Terrain Accidents Occurring Between 1990-1998," p. 25, 2003.

[8] S. Shappell, C. Detwiler, K. Holcomb, C. Hackworth, A. Boquet, and D. A. Wiegmann, "Human Error and Commercial Aviation Accidents: An Analysis Using the Human Factors Analysis and Classification System," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 49, no. 2, pp. 227–242, 4 2007.

[9] A. Agovic, H. Shan, and A. Banerjee, "Analyzing aviation safety reports: From topic modeling to scalable multi-label classification," in *Conference on Intelligent Data Understanding (CIDU)*, 2010, pp. 83–97.

[10] C. Pimm, C. Raynal, N. Tulechki, E. Hermann, G. Caudy, and L. Tanguy, "Natural Language Processing (NLP) tools for the analysis of incident and accident reports," in *International Conference on Human-Computer Interaction in Aerospace (HCI-Aero)*, Brussels, Belgium, 2012.

[11] N. Tulechki, "Natural language processing of incident and accident reports : application to risk management in civil aviation," Ph.D. dissertation, Université de Toulouse, 2015.

[12] L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, and C. Raynal, "Natural language processing for aviation safety reports: From classification to interactive analysis," *Computers in Industry*, vol. 78, pp. 80–95, 5 2016.

[13] S. D. Robinson, W. J. Irwin, T. K. Kelly, and X. O. Wu, "Application of machine learning to mapping primary causal factors in self reported safety narratives," *Safety Science*, vol. 75, pp. 118–129, 2015.

[14] S. Robinson, "Multi-Label Classification of Contributing Causal Factors in Self-Reported Safety Narratives," *Safety*, vol. 4, no. 3, p. 30, 2018.

[15] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[16] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*. New York, New York, USA: ACM Press, 1999, pp. 50–57.

[17] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, and J. B. Edu, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *ICLRWorkshop*, 1 2013.

[19] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13. USA: Curran Associates Inc., 2013, pp. 3111–3119.

[20] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[21] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[22] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[23] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, p. 1, 12 2015.

[24] B. McCann, J. Bradbury, C. Xiong, and R. Socher, "Learned in Translation: Contextualized Word Vectors," in *NIPS*, 7 2017.

[25] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," in *ACL. Association for Computational Linguistics*, Melbourne, 7 2018.

[26] "ASRS Database Online - Aviation Safety Reporting System." [Online]. Available: https://asrs.arc.nasa.gov/search/database.html

[27] "Model evaluation: quantifying the quality of predictions." [Online]. Available: http://scikit-learn.org/stable/modules/model_evaluation.html#multiclass-and-multilabel-classification

[28] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and Optimizing LSTM Language Models," *CoRR*, vol. abs/1708.02182, 2017.

[29] J. Howard and others, "The fastai deep learning library, v0.7.0," 2018. [Online]. Available: https://github.com/fastai/fastai

[30] A. Ergai, T. Cohen, J. Sharp, D. Wiegmann, A. Gramopadhye, and S. Shappell, "Assessment of the Human Factors Analysis and Classification System (HFACS): Intra-rater and inter-rater reliability," *Safety Science*, 2016.