5-1-2021

Comparing Human and Machine Learning Classification of Human Factors in Incident Reports From Aviation

Claas Tido Boesser

Florian Jentsch

COMPARING HUMAN AND MACHINE LEARNING CLASSIFICATION OF HUMAN FACTORS IN INCIDENT REPORTS FROM AVIATION

Claas Tido Boesser & Florian Jentsch
University of Central Florida
Orlando, FL

Machine learning algorithms have become popular tools for automated classification of text; however, performance of such algorithms varies and depends on several factors. We examined how a subjective labeling process based on a human factors taxonomy can influence human, as well as automated, classification of safety incident reports from aviation. In order to evaluate these challenges, we trained a machine learning classifier on a subset of 17,253 incident reports from the NASA Aviation Safety Reporting System using multi-label classification, and collected labels from six human annotators for a representative subset of 400 incident reports each, resulting in a total of 2,400 individual annotations. Results showed that, in general, reliability of human annotation for the set of incident reports selected in this study was comparatively low. Performance of machine learning annotation followed patterns of human agreement on labels. Suggestions on how to improve the data collection and labeling process are provided.

Continuous advances in computing power, in algorithms, as well as research in the fields of Artificial Intelligence and Machine Learning, have led to an increased application of these tools to Human Factors. What once were laborious tasks that had to be performed by humans are becoming increasingly automated. As such, an increasing number of studies are being conducted that seek to use a variety of computational methods for the analysis of incident reports with text narratives; studies are spanning across several industries, such as aviation, medicine, construction, and the railroad industry, among others. In the field of aviation safety, valuable insight into inflight incidents can be gleaned by examining narratives provided by personnel involved in flight operations that are reported under the condition of confidentiality (e.g., Dekker, 2014; Wiegmann & Shappell, 2003). Using such incident reporting data, researchers have used a variety of techniques, including the usage of topic modeling/data reduction algorithms to identify latent structures in the data, assessing report similarity, automatically labeling and classifying reports, and visualizing the results (e.g., Irwin et al., 2017; Kuhn, 2018; Robinson, 2016; Robinson et al., 2015; Tanguy et al., 2016).

Analyzing and categorizing data such as text narratives presents unique challenges. Along with the sheer volume of available narratives and their text form comes the challenge of extracting trends and information from unstructured data. One way to gain insight and, in turn, reduce the complexity of the data, is through the categorization of such data according to a taxonomy (e.g., Bailey, 1994; Tanguy et al., 2016; Wiegmann & Shappell, 2003). For aviation safety and incident reports, one such implementation is the human factors taxonomy consisting of 12 different labels that is being used in the public self-reporting database of aviation incidents known as the *Aviation Safety Reporting System* (ASRS; see Federal Aviation Administration [FAA], 2011, for a description of the program).

In this study, we compared human and machine learning classification of human factors categories in aviation incident reports from the ASRS database. In the process, we identified the challenges with regards to human and automated annotation, beginning with examining the pertinent characteristics of incident narratives and taxonomies, evaluating different ways of annotating the data, assessing whether some human factors constructs are easier to label reliably than others, all while discussing the implications of what is learned with regards to automatic classification. A main focus of this study was on evaluation of the viability of automated text classification given a subjective classification process. We studied (a) whether human annotators would be reliable and consistent in assigning the same labels to reports, when compared to one another and to the codes given by the experts at ASRS, and (b) whether an automated machine learning classifier could be trained to do this task at better than chance level and/or at a similar performance as human raters. Arguably, if a machine learning classifier does not perform better than chance, or when human annotation of a taxonomy is at the chance level, the reliability of the whole approach is in question.

Method

Using purposeful sampling, six annotators were recruited for this study. Three of the annotators were required to have at least a 4-year undergraduate or master's degree in Human Factors, or an associated discipline such as Psychology. They also had to have commercial flying experience or familiarity with 14 CFR Part 121 Air Carrier operations (we called these the domain plus classification, or "D+C experts"). The three other annotators did not have any formal schooling in Human Factors, but they were required to have commercial flying experience as active or former pilots of 14 CFR Part 121 Air Carrier operations (we called these the "D experts").

The human annotation of the ASRS narratives was followed by a qualitative and quantitative data analysis using machine learning and applying a mixture of statistical analyses from various disciplines in order to evaluate reliability of human annotation, machine learning performance, as well as the resulting interdependencies. In summary, this study consisted of the following steps:

1. Extract data from the ASRS database for the training of a machine learning classifier (17,253 incident reports and their associated human factors labels).
2. Generate a representative subset of the extracted data for the purpose of human annotation (400 incident reports to submit to human annotation).
3. Collect data from annotators including human factors labels for incident reports, confidence measures for selected labels and overall comments, if any.
4. Analyze inter-rater reliability (IRR) measures between the existing labels (referent labels), the D experts, and the D+C experts.
5. Split the 17,253 ASRS reports into a training and a test set using stratified sampling. Extract features. Train machine learning classifiers. Measure performance of machine learning classifiers using 10-fold cross-validation. Compare and contrast performance between different classifiers and between classifiers and human annotation.
6. Evaluate results.

Results

We set out to compare human and machine learning classification of human factors in aviation incident reports. One influences the other—classification is required in order to train a supervised machine learning model. Therefore, we also examined the interaction between human and machine learning classification. Hypotheses were based on some premises, mainly that (a) reliability in human classification is above chance level, (b) reliability depends on annotator and report characteristics, and (c) training a machine learning model can, to some extent, be beneficial for the task of analysis and classification of incident reports. Throughout this study, it became evident that there was considerable variability in the labeling of incident reports. As such, some hypotheses were supported, whereas others were not.

As hypothesized, we found that IRR was dependent on the label. Some labels of the taxonomy were more agreed upon than others, and in fact by a fairly large margin. Figure 1 shows agreement on labels based on Krippendorff's (2004) α.

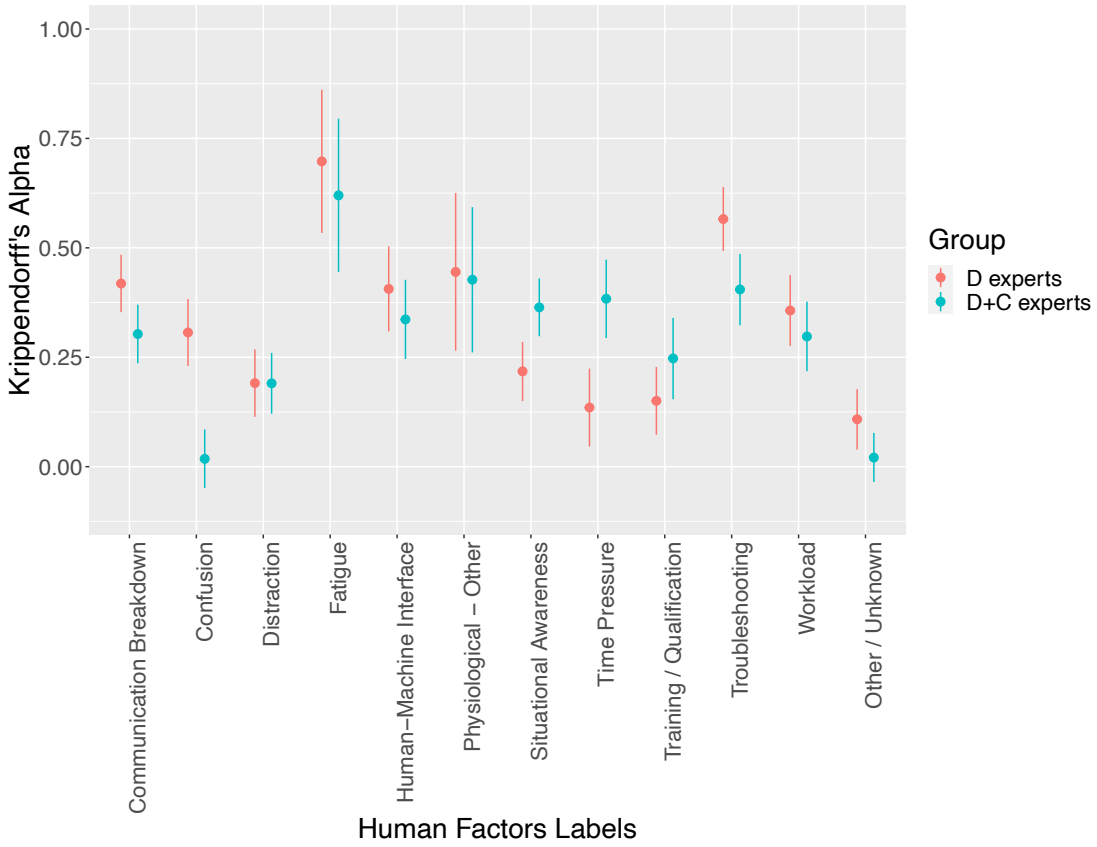

Figure 1. Krippendorff's α by label and group (D experts (left on each pair) vs. D+C experts (right on each pair)).

While, in general, agreement throughout the study seemed to be rather low, this is not necessarily unusual for the coding of raw incident reports. For example, Olsen and Shorrock (2010), as well as Olsen (2011) tested the reliability of the more widely researched HFACS

taxonomy—the original HFACS taxonomy in one study and a derivative of it in the other study—with conditions that closely resemble the research herein in the sense that there was no extensive training, and the incident report narratives were presented as the raw narratives to participants (as opposed to coding causal factors that were already abstracted from the reports). In their studies, agreement also highly varied depending on the specific HFACS category, but average percentage agreement at the category level was as low as 34.5% in Olsen and 39.9% in Olsen and Shorrock. This shows that the results presented herein are not necessarily unusually low when similar tasks are considered.

With regards to machine learning performance (see Table 1 for results), we found that, while human agreement and machine learning performance on labels did not exactly correlate with each other, there were some notable trends. For example, *Fatigue*, while not exhibiting a large prevalence in the dataset, stood out as one of the labels that were most agreed upon. *Fatigue* was also most reliably labeled by the machine learning classifier. As the prevalence of *Fatigue* was fairly low (only 5% of the original dataset contained the label), we followed up with a measure of separate agreement on the positive and negative class (see Feinstein & Cicchetti, 1990, as well as Cicchetti & Feinstein, 1990) and found a similar pattern, indicating that annotators were good at discerning when reports included fatigue but also discerning when they did not.

When examining the coefficients for the model, it also was evident that, for the label *Fatigue*, by far the largest predictor of the label was the occurrence of the actual word "fatigue." This poses the question of hand-coding rules versus machine learning. If only a few rules might lead to acceptable performance, why use machine learning to begin with? In fact, Tixier et al. (2016) achieved very good results with hand-coded rules for assigning attributes and outcomes to injury reports. However, they also noted that the process is tedious, labor-intense, heavily based on domain-knowledge and does not scale well to problems outside of the domain for which the rules were coded.

Table 1.
Precision, Recall, and F1-Scores for Individual Labels and Averaged Scores.

Labels	Precision	Recall	F1-Score	Support
Individual				
Fatigue	0.66	0.67	0.66	166
Communication Breakdown	0.60	0.73	0.65	1,267
Situational Awareness	0.62	0.59	0.60	2,017
Troubleshooting	0.44	0.82	0.57	649
Confusion	0.49	0.59	0.54	1,143
Physiological – Other	0.37	0.66	0.48	131
Human-Machine Interface	0.39	0.56	0.46	714
Workload	0.36	0.49	0.41	774
Distraction	0.37	0.45	0.40	903
Time Pressure	0.30	0.50	0.38	560
Training / Qualification	0.27	0.51	0.35	554
Other / Unknown	0.10	0.18	0.13	188

Note. Labels are presented in order of decreasing F1-score.

Discussion

Overall, there are clear challenges to be met in order to improve the annotation process both on the human and the machine learning sides, with one side influencing the other. DiMaggio (2015) wrote about the paradox that task performance of humans and a machine learning classifier often suffers at similar tasks. The research herein to an extent supports this statement. A straightforward categorization of "*Fatigue*", often based on the words, *fatigue*, *fatigued*, *tired*, or *sleep*, was more consistent than for concept labels such as "*Distraction*."

Other challenges that were discovered in the research herein illustrate the complexity of the problem, while also leading to valuable lessons learned. For example, evaluating performance on an imbalanced dataset is not straightforward as regular measures of accuracy are not appropriate for imbalanced data (for an overview, see Weiss, 2013, or Sahu et al., 2017). A similar challenge presented itself for the evaluation of IRR measures. As most IRR measures are sensitive to trait prevalence (e.g., Feinstein & Cicchetti, 1990; Gwet, 2008), imbalance in the data also needed to be accounted for with regards to measures of IRR.

In summary, there is promise in using ML with regards to fairly routine and simple categorizations. On the other hand, a basic ML algorithm, as used in this study, seemed to perform worse at anything that required more context and deeper analysis; but so seemed the humans. With that being said, categorizing narratives in accordance with a human factors taxonomy is an inherently subjective process. This leads to the conclusion that the labels that are provided either by the ASRS experts or by other annotators should always be seen as "a" categorization and not "the" categorization. Finally, recognizing the influence of narrative content as a major source of annotation variability is crucial to improving both the narrative, as well as the annotation. To improve the underlying quality of the reports, it is suggested to investigate, inter alia, automated cognitive aids based on the idea of semi-structured interview processes (see Crandall et al., 2006, as well as Wiegmann & von Thaden, 2003 for related ideas). For people involved in the creation and maintenance of incident databases, working together closely with human factors practitioners, as well as leveraging knowledge of the field of computer science should help to greatly improve incident reporting systems.

Acknowledgments

Participant payment for this study was partially funded by a 2018 *Human Factors and Ergonomics Society* Training Technical Group Student Grant Award. The paper is based on the doctoral dissertation of the primary author. The views of the research report reflect the views of the authors and not the views of the employers, the granting organization, or the University of Central Florida.

References

Bailey, K. D. (1994). *Typologies and taxonomies: An introduction to classification techniques*. Sage Publications.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*(6), 551–558.

Crandall, B., Klein, G., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*. The MIT Press.

Dekker, S. (2014). *The field guide to understanding "human error"* (3rd ed.). Ashgate Publishing Company.

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, *2*(2), 1–5.

Federal Aviation Administration [FAA]. (2011). AC 00-46E, Aviation safety reporting program.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48.

Irwin, W. J., Robinson, S. D., & Belt, S. M. (2017). Visualization of large-scale narrative data describing human error. *Human Factors, 59*(4), 520–534.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage Publications.

Kuhn, K. D. (2018). Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies*, *87*, 105–122.

Olsen, N. S. (2011). Coding ATC incident data using HFACS: Inter-coder consensus. *Safety Science, 49*, 1365–1370.

Olsen, N. S., & Shorrock, S. T. (2010). Evaluation of the HFACS-ADF safety classification system: Inter-coder consensus and intra-coder consistency. *Accident Analysis and Prevention, 42*, 437–444.

Robinson, S. D., Irwin, W. J., Kelly, T. K., & Wu, X. O. (2015). Application of machine learning to mapping primary causal factors in self reported safety narratives. *Safety Science, 75*, 118–129. https://doi.org/10.1016/j.ssci.2015.02.003

Robinson, S. D. (2016). Visual representation of safety narratives. *Safety Science*, *88*, 123–128.

Sahu, M., Mukhopadhyay, A., Szengel, A., & Zachow, S. (2017). Addressing multi-label imbalance problem of surgical tool detection using CNN. *International Journal of Computer Assisted Radiology and Surgery, 12*(6), 1013–1020.

Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., & Raynal, C. (2016). Natural language processing for aviation safety reports: From classification to interactive analysis. *Computers in Industry*, *78*, 80–95.

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*, *62*, 45–56.

Weiss, G. M. (2013). Foundations of imbalanced learning. In H. He, & Y. Ma (Eds.), *Imbalanced Learning.* IEEE Press.

Wiegmann, D. A., & Shappell, S. (2003). *A human error approach to aviation accident analysis: The Human Factors Analysis and Classification System*. Burlington, Ashgate.

Wiegmann, D. A., & von Thaden, T. L. (2003). Using schematic aids to improve recall in incident reporting: the critical event reporting tool (CERT). *The International Journal of Aviation Psychology*, *13*(2), 153-171.