Subjective Machines: Probabilistic Risk Assessment based on Deep Learning of Soft Information

Abstract

For several years machine learning methods have been proposed for risk classification. Whilst machine learning methods have also been used for failure diagnosis and condition monitoring, to the best of our knowledge, these methods have not been used for probabilistic risk assessment. Probabilistic risk assessment is a subjective process. The problem of how well machine learning methods can emulate expert judgements is challenging. Expert judgements are based on mental shortcuts, heuristics, which are susceptible to biases. This paper presents a process for developing natural language-based probabilistic risk assessment models, applying deep learning algorithms to emulate experts' quantified risk estimates. This allows the risk analyst to obtain an apriori risk assessment when there is limited information in the form of text and numeric data. Universal Sentence Embedding (USE) with Gradient Boosting Regression (GBR) trees trained over limited structured data presented the most promising results. When we apply these models' outputs to generate survival distributions for autonomous systems' likelihood of loss with distance, we observe that for open water and ice shelf operating environments, the differences between the survival distributions generated by the machine learning algorithm and those generated by the experts are not statistically significant.

1. Introduction

Substantial developments in machine learning techniques, enhancements in computer performance, and better data collection processes have led to an increased desire and ability to use machine learning methods for risk quantification (Paltrinieri et al., 2019). Machine learning algorithms have been used to support risk analysis in applications where there is a substantial amount of data. Areas of successful application are as diverse as, for example, engineering machine anomaly detection (Garcia, 2013, Hodge and Austin, 2004) and financial risk, e.g., credit risk and default risk (Lessmann et al., 2015). That said, some risk domains are still heavily dependent on expert judgements.

To this date, machine learning methods have not been used for probability risk assessment. These probabilities include, for example, that of a nuclear reactor failure (van Steen, 1992) or major civil catastrophes (Bigün, 1995). This has not occurred for two reasons. First, probabilistic risk assessment is expert knowledge-dependent, as this is knowledge of the subject area and structured knowledge about the problem. One type of expert knowledge can be, for example, causal relations between factors (Aven, 2016). Second, the expert's estimate of the likelihood of incident occurrence is subjective and susceptible to bias (Kahneman and Tversky, 1972, Morris, 1977, Tversky and Kahneman, 1974). Machine learning algorithms can capture randomness well, but to the best of our knowledge, subjectivity has not been addressed in machine learning approaches for risk assessment. Approaches to mimic expert assessment in critical applications rely on knowledge-based systems that capture cause and effect phenomena using methods such as Bayesian Belief Networks (Hänninen and Kujala, 2012, Kabir et al., 2015, Sigurdsson et al., 2001). One of the benefits of Bayesian belief networks is that they enable supervised learning as data become available (de Campos and Castellano, 2007). Our approach improves upon the use of this technique in several ways. First, Bayesian Networks require a dependency

graph to be estimated for the causality paths to be established. In text-based models, this process can be prohibitively expensive computationally, or would require strong assumptions to be made about the distributions in the data. The models we use in this work go beyond these needs by not needing any assumptions over the dependency of the decision, instead leaving the discovery of the relevant parts of the evaluations (the "dependencies" on a text framework) to be automatically generated by the self-attention mechanisms of the neural network. This explicit vs implicit representation of the uncertainty in decision-making can be useful when there is limited (or no) knowledge of the potential dependency paths. A neural network model is then able to measure this uncertainty from past behaviours instead of "hard-coding" them, as a Bayesian Network requires them. This is a more flexible approach, albeit one more uncertain. Second, an empirical model such as the one we present can be used to create future dependency graphs for potential Bayesian models, trading off the background knowledge necessary for the success of the Bayesian network with the necessity of a larger dataset where the empirical Neural Network model can operate. This trade-off is, of course, application specific, but it is expected in data-rich environments such as the ones we discuss in this work.

Natural language machine learning methods have been proposed to give insight into sentiments presented in narratives. Initially developed in the 1970s, natural language techniques are now used in web applications for profiling. Sentiment analysis techniques used in natural language machine learning have been proposed to support risk analysis. However, to date, sentiment analysis has not been used for probabilistic risk assessment.

For many years, researchers have investigated whether artificial intelligence (AI) systems can replace human reasoning. The answer to this question has consistently been that AI-based systems can outperform humans for certain specific applications, but for other applications, AI systems cannot achieve human-level performance. For example, for certain games such as

chess, AI systems have achieved superhuman performance for more than a decade (AI Performance, 2020) while, in general, conversational AI systems (chatbots) run into severe issues when used for customer engagement (Adam et al., 2020). The use of machines to reciprocate expert assessments in the context of AUV has never been explored to the best of our knowledge. This paper proposes a methodology for developing sentiment analysis methods for conducting risk assessments otherwise carried out by experts.

Our paper's contributions are three-fold: First, we present a general methodology to construct such models using the latest advances in deep learning methods combined with subjective statistical survival estimator. Second, we present a methodology to treat and prepare the data for such models, pooling the opinions of the experts. Finally, we perform a detailed analysis and validation of these models' outputs to prove their effectiveness and discuss their validity beyond statistical accuracy, including context learning.

This paper is organised as follows. Section 2 presents related work in sentiment modelling methods for risk analysis. Section 3 presents the proposed methodology and the data used for testing. Section 4 presents the results. The discussions and conclusions are presented in sections 5 and 6, respectively.

2. Related work

Text mining and sentiment analysis are two different tasks under the umbrella of natural language processing. Text mining is a technique for knowledge retrieval of unstructured text. It is used in more visible ways for internet applications (Mohammad and Turney, 2010). Technologies in the text mining realm include information extraction, topic tracking, summarisation, categorisation clustering, concept linkage, information visualisation, and question answering (Fan et al., 2006).

Text mining has been applied in many industry sectors. Holton (2009) applied text mining techniques to identify disgruntled employees by collecting data from Vault.com and

Yahoo.com. The author used a total of 44 disgruntled reports and 44 non-disgruntled reports, plus 10 disgruntled. The author applied a Naïve Bayes (NB) model to classify disgruntled and non-disgruntled reports (Domingos and Pazzani, 1996). The NB model performance was satisfactory, with a recall of 90% for non-disgruntled and 87% for disgruntled. The method proposed by Holton (2009) is efficient as a tool to inform employers of potential disgruntling in the organisation. That said, disgruntling on its own does not present a fraud risk. In general terms, risk is defined by the triplet of *hazard*, *likelihood* and *consequence* (Kaplan and Garrick, 1981). Identifying whether or not someone is disgruntled presents a hazard; alone, however, it does not define the risk of fraud.

In finance, Groth and Muntermann (2011) proposed a text mining approach to make predictions of asset volatility levels for the time period that immediately follows the publication of corporate disclosures. The authors explored 423 disclosure reports and volatility risks immediately after their submission. They tested Naïve Bayes, k-Nearest Neighbour (kNN), Neural Networks (NNet) and Support Vector Machine (SVM)(Vapnik, 1995). In this and other research, the risk quantity is not a probability judgement provided by an expert. Instead, the risk is calculated using a set of rules. For example, in the case of Groth and Muntermann (2011), volatility is calculated using an established rule. In other applications of text mining, the risk calculation is usually a separate exercise. For example, Lee and Yi (2017) quantified risk as a factor of the project bid cost while Son and Lee (2019) defined risk as a factor of project delay. Text mining has also been used to identify risk events from accident reports (Sarkar et al., 2016).

Sentiment analysis is a subset of text mining. Sentiment analysis attempts to measure the inclination of people's opinions based on natural language processing. This technique alone cannot estimate probabilistic risk. Thus far, the application of sentiment analysis has been

focused on predicting judgements in binary decisions (Medvedeva et al., 2019, Liu and Chen, 2018) or the occurrence of discrete events (Yadav et al., 2019).

Probabilistic risk assessment (PRA) is an approach first introduced in the 1970s to capture the uncertainty in mathematical models used for nuclear reactor risk analysis (Apostolakis, 1990). This method was later adopted in other sectors of the nuclear industry – e.g., nuclear waste risk estimation (Bonano et al., 1990). Nowadays, PRA is used for risk analysis of complex systems for which there is a modicum of data, which alone cannot be used to characterise the risk fully.

As more data from PRA are made available it becomes possible to use machine learning and sentiment analysis to model the effect of context on the risk estimation. A framework based on these methods can be used to provide probabilistic risk assessments of future scenarios, not considered by the experts. The methodology for developing this framework is presented in the next section.

3. Methodology

Expert judgement elicitation is a process developed by mathematicians and social scientists to elicit judgements from experts with respect to uncertain quantities (Cooke and Goossens, 2004, O'Hagan et al., 2006, Otway and von Winterfeldt, 1992) that takes into consideration that experts are susceptible to motivational and subconscious bias (Merkhofer, 1987, Tversky and Kahneman, 1974).

In this section, we explain how expert judgements are elicited for autonomous systems risk assessment and the required assumptions to implement a text mining solution to mimic expert judgements. A flowchart of the methodology is presented in Figure 1.

3.1 Expert judgement data

The data used for the analysis were collected from two formal expert judgement elicitation processes. These were conducted for quantifying the risk of autonomous

underwater vehicle (AUV) loss in extreme environments, open water, coastal waters, sea ice, and ice shelves. The experts were given information about the AUVs, the operating environment, and the fault description (Brito et al 2010). We present three samples of these fault descriptions below:

Sample 1:

Aborted after 4 minutes post dive, due to network failure. Logger data showed long gaps, up to 60s, across all data from all nodes, suggesting logger problem.

Sample 2:

Jack-in-the-box recovery float came out, wrapping its line around the propeller, jamming it, and stopping the mission. Caused severe problems in recovery, some damage to upper rudder frame, sub-frame and GPS antenna. Required boat to be launched.

Sample 3:

Pre-launch, potential short circuit in motor controller that could stop motor.

The experts were asked to estimate the probability of fault leading to loss in each of the four environments P(L|F,E).

The first expert judgement elicitation was conducted for building the risk model for the *Autosub 3* AUV vehicle deployment in the Pine Island glacier in Antarctica (Brito et al., 2008). This expert judgement was conducted using the Otway and Winterfeldt method (Otway and von Winterfeldt, 1992). Experts have provided individual estimates for the probability of AUV loss given a fault in a given environment P(L|F,E), as well as a weight w_i (from 1-5) for capturing their confidence in the estimate (a weight of 1 if not confident and a weight of 5 if very confident). Eight experts took part in the expert judgement elicitation; these were AS, BF, CJ, CW, DY, MM, RM and TC. A total of 63 faults were assessed by the experts. Of the 2016 expected individual estimations, 1141 individual expert judgements were collected.

The second source of data is from the expert judgement conducted for estimating the risk of *Remus 100* loss (Griffiths et al., 2009). This expert judgement elicitation was conducted using the SHELF-R method (Garthwaite et al., 2005, O'Hagan et al., 2006). Five experts took part in this expert judgement elicitation –namely, AS, CW, RM, SM and TB. The experts were asked to estimate the probability density distribution of the probability of fault leading to loss in a given environment. The experts were provided with the lower bound, upper bound, median, lower quartile and upper quartile to describe the distribution. A total of 504 estimates were obtained. Three of the experts took part in both expert judgements elicitation – namely, AS, RM and CW.

For both expert judgement elicitations, the experts were independent of the organisation for which the probabilistic risk assessment was conducted, and had received adequate training prior to the elicitation. Also, for both elicitations, the expert judgements were aggregated using the linear weighted pool (Winkler, 1968).

Following a review of expert judgements, the facilitator concluded that experts' assessments had be grouped into *optimists* and *pessimists*. The concept of optimists and pessimists is different from that presented in Goodwin et al. (2019) where experts provided their optimists' and pessimists' forecasts. Here one expert is considered an optimist if she uses predominantly low probabilities, and the narrative suggests that the expert gives more weight to positive consequences of the hazard event than to negative consequences. One expert is considered a pessimist if she uses predominately high probability ranges, and the analysis of the narrative suggests that she gives more weight to negative consequences of the hazard event. The purpose of aggregating the judgments into these two models is to give to the decision maker two decision models to select from: one optimist and one pessimist. Upon analysis of the narratives provided by the experts and the judgments, the decision maker chooses which model is more aligned with her beliefs.

In both elicitations the facilitator analysed the narrative provided by the experts to justify their likelihood estimation. The facilitators concluded that when estimating the likelihood of a fault leading to loss, all experts identified the same secondary risks but some experts would take an optimist's or a pessimist's view of the impact of these risks; for example, if a fault led the AUV to come to surface without power, unable to communicate and unable to navigate. One secondary risk that may emerge here is that the AUV may be dragged to the coast and crushed against the rocks. Whilst this was a possibility, some experts would consider that, following the same fault, the AUV would be dragged to a sandy beach and safely returned to the users (Brito et al., 2008, p.14). When this was the case, the probability for loss given a fault in a given environment (P(L|F,E)) of an optimist was significantly lower than that given by the pessimist. Linear weighted mathematical aggregation in this case would censor the estimation given by the optimist. In light of the large volume of assessments, the facilitators used cumulative plots to visualise the tendency for experts to use higher or lower probability values. For sea ice and ice shelf, in particular, the facilitators identified different shapes of the cumulative distributions, depending on whether the experts had the tendency to be optimists or pessimists (Brito et al., 2008, p. 21). The same process was followed to aggregate the judgments from the *Remus 100* AUV expert judgement elicitation (Griffiths et al., 2009). Based on the narrative analysis and cumulative plot analysis, the facilitators identified groups of optimists and pessimists. For more detail we direct the reader to Brito et al. (2010) and Griffiths et al. (2009) who present the dataset and the results of the risk analysis conducted on the *Autosub 3* and the *Remus 100,* respectively.

Brito and Chang (2018) analysed the risk estimates provided by optimists and pessimists used to build the risk model for a Hybrid AUV, *Neureus UI*, owned and operated by Woods Hole Oceanographic Institute. Data collected during actual missions allowed the researchers to compare the expert risk predictions from optimists and pessimists to the actual mission

data. The researchers concluded that the model developed based on the pessimists' assessments was more accurate than the model created based on the optimists' assessments.

In this paper we do not attempt to evaluate the reliability of different groups of experts. Here we simply use the risk models developed based on the expert judgement elitations for *Autosub 3* and *Remus 100* to train the machine learning algorithm.

3.2 Data preparation

The raw dataset contained 252 faults (63 *Autosub 3* and 183 *Remus 100*); however, the number of faults had to be significantly reduced due to acceptability criteria. The following criteria were applied: First, only faults for which all experts provided assessments were considered and, second, faults considered had to have a confidence weight of at least three. As a result of these assumptions, the number of single faults was reduced to 100.

With 100 unique faults, four environments and two expert types (optimistic and pessimistic), the final dataset size is 100*4*2 = 800. Table 1 displays the features used in the machine learning models.

Table 1. Features used in the machine learning models.

Feature	**Classification**	**Type**	**Description**
Distance	Structured Features (Core Data)	Continuous	The measure in metres that the fault occurred from the mission start point
Condition		Categorical	One of Open Water, Coastal, Sea Ice or Shelf Ice
AUV Model		Binary	AUV model type; Remus or Autosub3
Expert Type		Binary	Reviewer classification; Optimistic or Pessimistic
Text Description	Text	String	Short description of the fault produced by mission crew

For the structured features, a standard approach is applied. For the single continuous variable, 'distance', a Min/Max procedure rescales the variable between 0 and 1. This ensures

that all input data for the machine learning algorithms are within the same scale, which is particularly pertinent for the GLM model. As the machine learning models cannot take text strings as input, the categorical data ‘Expert Type’, ‘AUV model’ and ‘Condition’ are One-Hot encoded, producing an output binary vector, e.g., ‘Condition=Open Water’ is transformed to [1,0,0] while ‘Condition=Coastal’ becomes [0,1,0].

As with the categorical data, it is a requirement for the machine learning models that the text be represented as a numerical vector. Unlike the categorical representation, however, One-Hot encoding is not viable due to the text's extreme cardinality – i.e. the total number of words. Instead, we obtain vectors representing the text using four methods that vary in complexity; LSA, USE, ELMO and BERT.

The simplest representation applied is Latent Semantic Analysis (LSA; Landauer et al., 1998). LSA might be considered an early approach to Natural Language Processing (NLP) although it is still widely used. The method derives ‘concept’ features using predefined statistical techniques and assumes word co-occurrence across texts that share meaning, independent of sentence structure. Although simple, it requires the heaviest pre-processing to reduce the variability of the text. Pre-processing includes the removal of stop words (e.g., it, there, was, so) and the stemming of words to remove suffixes (e.g., breaking -> break). There are two stages to vectorising the processed text. First, a term frequency-inverse document frequency (TF-IDF) matrix is derived, producing a vector the length of the entire corpus vocabulary for each text description. A Truncated Singular Value Decomposition (SVD) approach is then used to reduce the TF-IDF matrix's size and complexity. Fifty latent concepts are selected using the elbow method; the ‘elbow’ is the point at which n+1 components contribute diminishing returns to the explained variance. This value was found to meet this criterion and also provide additional redundancy as there is a further process for screening redundant variables.

The field of NLP has seen significant advances beyond approaches such as LSA from the field of Deep Learning (Goodfellow et al., 2016). Word-level embeddings obtained from a model trained on a universal language task, e.g., missing word prediction can be used on a downstream task (Goldberg and Levy, 2014). More recent Deep Learning approaches are capable of encoding entire text sequences; thus, they can also account for positional and contextual structures in the text. Three models which utilise such advances are used to obtain the embeddings: BERT (Devlin et al., 2018), ELMO (Peters et al., 2018) and USE (Cer et al., 2018). The use of these embeddings has an added benefit over the LSA approach, which we leverage in this work — they utilise transfer learning. Transfer learning in the context of language modelling is an approach in which the Deep Learning models are pre-trained over an extremely large corpus of text, usually with an unsupervised task; i.e. not requiring labels. This pre-training allows the model to learn complex but general language structures reducing the requirement for further training and is particularly pertinent for our dataset, given the limited number of text fault descriptions.

The Universal Sentence Encoder (USE) model, proposed by Cer et al. (2018), is distributed with several versions on which we use the Deep Averaging Network (DAN) implementation. The model operates by averaging word embeddings and word bi-gram embeddings, subsequently used as input to a Deep Neural Network (DNN). So that the model learns general representations from the text, it is trained in a multi-task setting to predict several supervised and unsupervised learning tasks. We then utilise the pre-trained model encodings with the fault description text, forming the vector inputs to our models.

ELMO (Embeddings from Language Models), proposed by Peters et al. (2018), extends earlier work utilising word embeddings by producing ‘contextual’ word embeddings. These embeddings look at the fixed absolute representation of a word and its relation to surrounding words in a sentence or statement. The embeddings are produced by training a Bidirectional

Long short-term memory (bi-LSTM), a specific type of Deep Learning model, to predict the next word in a sentence over a large text corpus. Once this unsupervised training has occurred, the hidden states of the bi-LSTM can be unrolled and aggregated to produce an average vector representation for a given text statement.

Bidirectional Encoder Representations from Transformers, or BERT (Devlin et al., 2018), builds upon ELMO. Amongst the most notable differences are the use of a powerful Transformer model rather than a bi-LSTM encoder. Additionally, the model is trained using two unsupervised tasks, including masked (missing) word prediction and next sentence prediction. Like ELMO, from the pre-trained model, we extract the pooled representation for each text description.

All three of the approaches require little pre-processing beyond tokenisation. It is usually desirable to undertake a further fine-tuning stage of the pre-trained models on the downstream task. However, given the small number of cases, the output of the models we propose here can be used as raw feature extractors with no further training. The final embeddings per case are sizes 512, 1024,1024 for the USE, ELMO, and BERT, respectively.

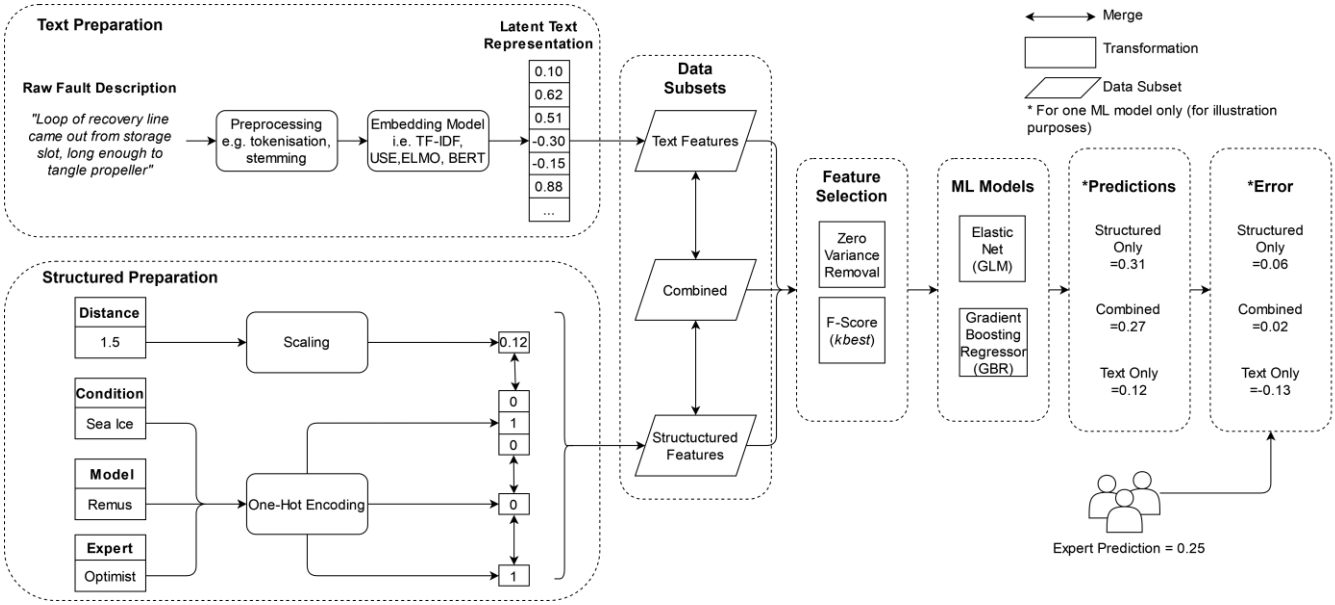

Figure 1: Methodology for developing machine learning algorithms for predicting expert estimations of the risk of losing an autonomous underwater vehicle.

3.3 Regression task

Two standard machine learning model types are applied to the data providing differing levels of complexity; the Generalised Linear Model (GLM) and the Gradient Boosting Regressor (GBR). Both models deviate slightly from their standard implementation as they are Beta-constrained to ensure predicted values between 0 and 1 (Ferrari and Cribari-Neto, 2004). The GBR is the most complex of the models and capable of learning non-linear patterns and complex variable interactions. Conversely, the GLM provides a simpler model against which the GBR can be benchmarked. While there are many other supervised Machine Learning algorithms, these two implementations give a good range by which we can assess the viability of our 'subjective machines'.

In this section, the details of the model types and approach to hyperparameter optimisation are described.

3.3.1 Model Training Strategy

In total, there are 18 models covering the data subsets and model types. Although each model is independently trained, a common approach is applied to the implementation.

Before training each of the models, an initial round of feature screening is applied. First, features with zero variance are removed. Following this, univariate feature screening is applied using the regression F-Score. Although both models are theoretically capable of screening redundant variables, given the relatively few number of faults and a large number of features, this ensures that the noise in the data is minimised. Based on the F-score, the *kbest* features are selected from the initial subset. The *kbest* parameter is optimised on the search space, ranging from 1 up to the given subset's total number of features.

Bayesian cross-validation optimisation is applied to select the *kbest* parameter and the model-specific parameters (Head et al., 2018). With this approach, a generous search space is initially provided across the parameter space with 50 seed models tested across 10-folds of the data. The parameters are subsequently optimised over 50 iterations. Fifty iterations were selected to provide sufficient redundancy to converge performance on the most complex model/embedding condition.

3.3.2 Generalised Linear Model (GLM)

The Elastic Net variation of the GLM is applied to the dataset, providing a compromise between the Lasso (L1) and Ridge (L2) regularisation approaches (Zou and Hastie, 2005). With Elastic Net, the *L1 Ratio* parameter can be optimised, allowing the flexibility to select either L1 or L2 penalisation, or a mixture of both. This flexibility is particularly important as the number of features in the dataset varies significantly depending on the feature subset. Both the *L1 Ratio* and the *alpha* (the extent of the penalisation) are optimised using the Bayesian approach.

3.3.3 Gradient Boosting Regressor (GBR)

Gradient Boosting is a tree-based ensemble approach that iteratively builds weak learners that seek to minimise the residual, eventually developing a single strong learner (Friedman, 2001). Gradient Boosting is both non-linear and efficient at feature selection; therefore, the model is expected to outperform the GLM. Bayesian optimisation is applied to the following parameters*: learning rate, n estimators, min samples split, max depth* and *max features*.

3.4 Experimental design

This paper seeks to assess both the machine learning models' predictive capabilities and the predictive capacity of the text embeddings. As such, the models are trained on the structured data only, on the text only, and on the combined structured and text inputs. Thus, the data subsets are as follows:

1. Core Data (structured data only)
2. LSA (with/without structured features)
3. USE (with/without structured features)
4. ELMO (with/without structured features)
5. BERT (with/without structured features)

Each unique fault in the dataset is represented by four conditions (Open Water, Coastal, Sea Ice, Shelf Ice) and two expert types (Optimist/Pessimist), producing eight cases per fault in total.

It is standard procedure in machine learning to split the input data into a training set, which is used to train the model, and a test set used to evaluate performance. Since there are just 100 unique faults, a single hold-out sample would be unlikely to produce a robust model performance estimate. Accordingly, we apply cross-validation which uses multiple train/test sets splits, iteratively retraining the models and evaluating performance. Two types of cross-validation are applied: Leave-One-Group-Out (LOGO) and Leave-One-Out (LOO). The

LOGO approach iteratively tests a single fault, therefore removing eight cases per iteration. The LOO approach, on the other hand, removes just an individual case per iteration. LOGO validation is more challenging compared to LOO as all inputs are unseen by the model. The real-world performance is likely to fall somewhere between the two validation approaches.These valiadation strategies are presented in Figure 2.

Furthermore, in operational terms, each represents a different scenario. LOGO is broadly equivalent to predicting an unseen fault with no prior expert assessment. LOO more closely represents performance given some known information about a fault that has occurred; for example, moderating a single expert's assessment.

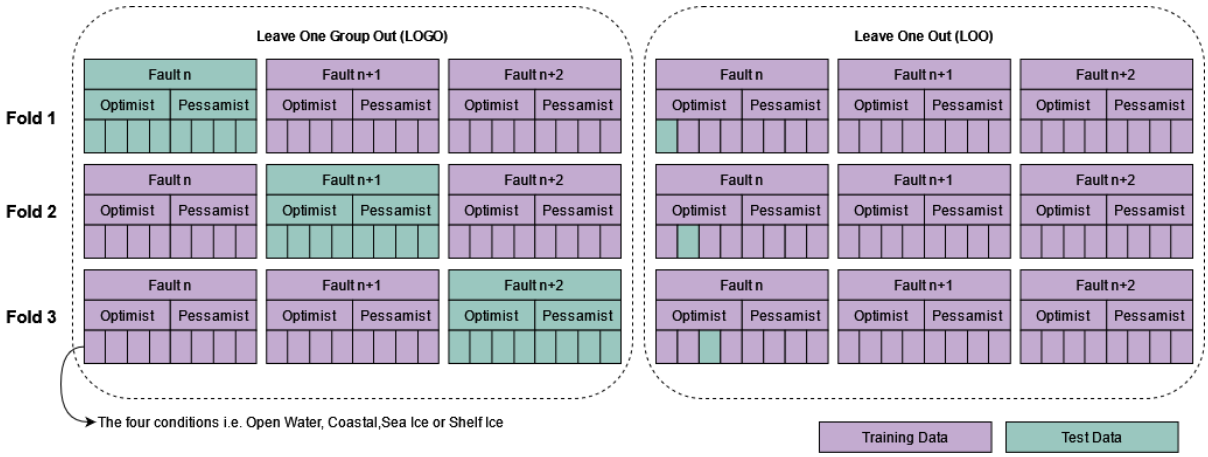

Figure 2: Leave-One-Group-Out (LOGO) and Leave-One-Out (LOO) performance evaluation strategies.

3.5 Subjective Survival Estimator for mission Risk Analysis

Survival modelling methods are a family of analytical models that provide inference about survival data. In conventional survival modelling the observation dataset *x* contains censored and non-censored data (Kaplan and Meier, 1958). Each element in the dataset is a time observation of when a fault has occurred (non-censored observation) or when a mission has been completed without a fault. If a mission was completed without a fault this observation is deemed censored. The subjective Kaplan Meier (K-M) survival estimator uses a similar type of data to conventional survival models. The subjective K-M survival estimator

uses the same censored data as the conventional survival estimator. However, for the subjective K-M survival estimator, the non-censored data is the likelihood of a fault leading to AUV loss in a given environment. The likelihood of a given fault in a given environment leading to AUV loss is estimated by experts. The subjective Kaplan Meier survival estimator, $\hat{S}$, for quantifying the probability of survival with distance x is presented in Eq. (1), below.

$$\hat{S}(x) = \prod_{x_i<x}\left(1 - \left(\frac{1}{n_i}\right)P(L|F_i,E)\right) \quad (1)$$

A single AUV mission (either failed or successful) is considered as an event. All events are assigned the decreasing index n_i according to the mission distance at which it ended (regardless of the outcome). For each fault, F_i, a group of experts is asked to agree on the probability of fault leading to AUV loss, given that it is operating in a target environment E. This is the probability of failure, it is a conditional probability,$P(L|F_i,E)$.

4. Results

In this section we present the analysis of subjective machine risk prediction. The results obtained from the subjective machine risk prediction are compared to the pooled human experts' risk prediction.

4.1 Error Analysis

Figure 3 presents the predictive performance in terms of the root-mean-square error (RMSE) metric. Each column of grids is represented by the model (GLM & GBR), while each row of grids is the type of validation used (LOGO & LOO). Each grid consists of nine result tiles, with the columns reflecting the embedding type while rows reflect the data subset. Blank spaces are where the cross-section is mutually exclusive – i.e. there is no embedding when only the core structured data are used. Values in bold format represent the best performing model in each grid.

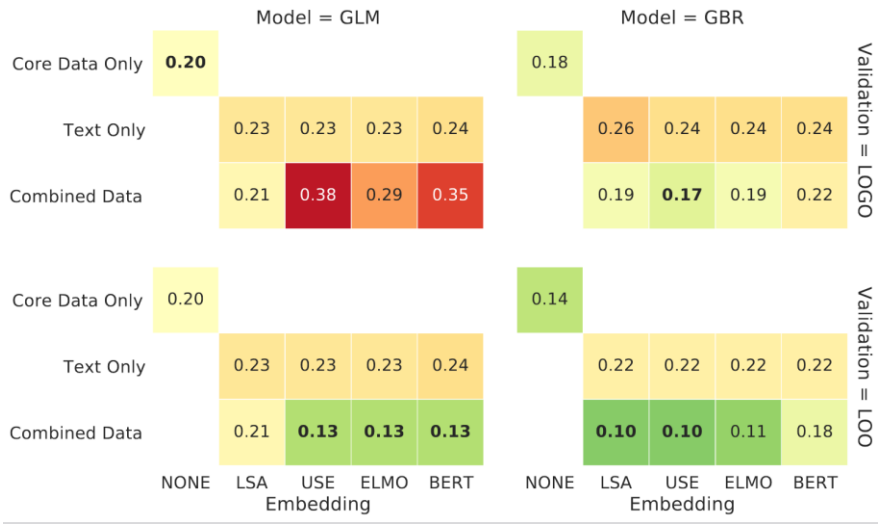

Figure 3: Predictive performance (RMSE)

Considering first the performance on the core data only, the results range between 0.10 and 0. 38, with the GBR model markedly outperforming the GLM. Superior GBR results suggest a non-linearity in this data subset that is better captured by the GBR model. The RMSE scores also appear reasonable, suggesting that, even with some basic information about the mission, fair performance can be gained using machine learning approaches.

When reviewing the text-only performance, across none of the data subsets, models or embeddings does the text demonstrate predictive performance. The best performing models produce an RMSE score of 0.23 – arguably noise. However, the combined models' results using both the text and core structured data tell a different story with substantial increases in performance over the core data models present. This is particularly true of the USE embedding which demonstrates high RMSE scores ranging between 0.10 and 0.17, except the GLM LOGO condition. The performance of the GBR with USE using LOGO validation is of particular significance. Under this condition, the combined model provides an uplift of 0.1 (5.6%) over the core data subset.

Performance with LOO validation only might suggest that the embedding provides enough noise to overfit to a particular fault. Therefore, the observed performance with LOGO

– holding out an entire fault – instead suggests that the text does not contribute noise but valuable new information not included in the core data. The text information enriches structured data when it is used simultaneously. Performance increases over the core results are also present for the LSA, ELMO and BERT embeddings. However, this tends to be restricted to LOGO validation.

Additionally, the GBR model notably outperforms the GLM for the combined input subset, which is somewhat expected as the GBR has a higher capacity to capture the non-linearity and feature interactions in the data. Moreover, it might also explain why the text does not yield a positive result alone but, combined with the core data, a positive result is observed. It could be argued that the text simply presents too much noise, particularly so for a small dataset such as this one. However, the structured data allow the model to point to the embedding space's relevant areas, a task better suited to the GBR model.

USE embeddings outperforming the more advanced BERT and ELMO models might be considered somewhat surprising. This could be explained by the fact that these approaches may perform better on larger datasets and that, with such datasets, the models will undergo further fine-tuning. Therefore, on our dataset, ELMO – and particularly BERT – are not as effective as raw feature extractors compared to USE for a small dataset such as this one.

4.2 Model explanation

Local surrogate models are interpretable models that are used to explain individual predictions of machine learning models. Local interpretable model-agnostic explanations (LIME) is an implementation of a local surrogate model (Ribeiro et al., 2016). We have chosen to use the LIME method to assess the impact of different words on the GBM model predictions. This is one of the most widely used techniques for this task. LIME enters variations in the data and measures the impact on the machine learning model. We have adopted a LIME library implemented in Python, which iteratively drops each word from

single-fault descriptions and monitors the impact of each word's inclusion. The results of the LIME analysis are presented in Figure 4. These show the top 30 words for high-risk estimation and the top 30 words for low-risk estimation. On the x-axis is each word, while the y-axis is the mean impact of the word inclusion over all fault descriptions when it was included. For example, a positive impact score of 0.05 for a given word would indicate that the model produced a higher risk of loss prediction (by 0.05 or 5%).

As shown in Figure 4 (top), the words *science*, *pier*, *pinger*, *lock* and *leaking* are the five most significant words in the high-risk estimation. In AUV risk, these words are significant; *science* is related to the science bay of the AUV. A leak in the science bay may lead to AUV loss. It may change the buoyancy of the AUV rendering it unable to come to surface or it may cause a significant electrical fault. The word *pier* concerns the operation of *Remus 100*. This vehicle operates near coastal areas. One of the key risks is a collision with a pier. The words *pinger* and *lock* are associated with the ability to locate the AUV. These are also associated with a high risk of loss. The word *leaking* is associated with high risk for the reasons mentioned earlier. The Autosub 2 AUV is thought to have been lost due to a leak (Strut, 2006).

Looking at Figure 4 (bottom), the top five most significant words that can explain low-risk estimation are *making, rpm, sound, sensors,* and *gain*. These words are associated with concerns over navigation performance or science data-gathering. These words can explain

low-risk estimation because they relate to mission performance optimisation and not to an

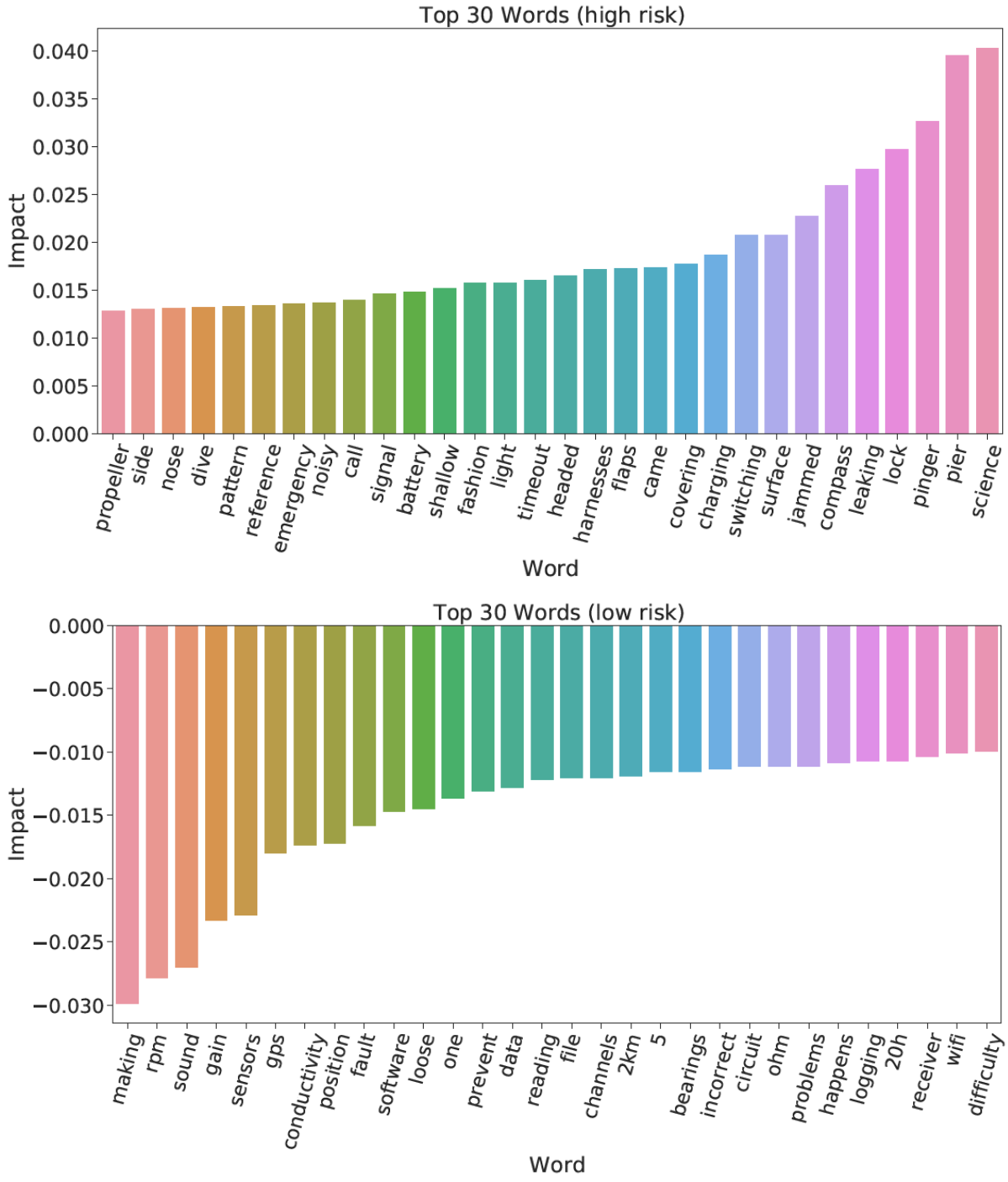

Figure 4: Output from LIME analysis. Top: Most significant 30 words for high-risk estimation.

Bottom: Most significant 30 words for low-risk estimation.

4.3 Single failure analysis

In subsection 4.1 we showed that the the GBR algorithm using the USE embendings provide the minimum RMSE and therefore it is the most suitable algorithm for this application. But what does this mean in terms of individual failure risk analysis? If we consider a single fault, for example fault 392_1_1, is one of the top five most critical faults in the Autosub3 failure analysis. The fault is described as follows: GPS antenna flooded. No fix at the end point of the mission. AUV ended up 700m North and 25m East of the expected position. The assessments from the optimisits and pessimist groups are presented in annex F of Brito et al (2008).

For the open water, considering the experts aggregation using the linear pool, the probability of loss given fault 392_1_ 1 is 0.00305 and 0.0346, for the optimist and pessimist models respectively. For the ice shelf environment the probability of loss given fault 392_1_1 is 0.105 and 0.222 for the optimist and pessimist models respectively.

The GBR USE model considering the LOGO validation, for open water predicts the probability of loss given fault 382_1_1 as 0.00234 and 0.036, using the optimisitic and pessimisitic model respectively. The GBR with the USE embedings provide slightly lower risk estimate using the optimisitic model and slightly higher risk estimate using the pessimisitic model. Considering the same model, with the LOO validation, the probability of loss given fault 392_1_1 is 0.0185 and 0.0427, for the optimist and pessimisit model respectively. Here both estimates are slightly higher than the experts aggregated judgment. As the results show the differences between the GBR model using USE embendings and the expert aggregated judgements are small and within the same order of magnitude.

For ice shelf environment the probability of fault leading to loss considering the LOGO validation is 0.065 and 0.192 for the optimist and pessimist models respectively.The LOO validated model estimates the risk as 0.0185 and 0.4426, using the optimisit model and the

pessimisit model respectively. Here the risk is lower for the optimisit model and higher for the pessimist model.

This example, shows that regardless of the type of the validation used the GBR model with USE embebdings provides a good prediction of the experts estimations. Considering narrative, the word GPS is the only word which has a significant weight on the risk estimation. The LIME results in Figure 4 show that the word GPS has an impact on low risk estimates.

4.4 Subjective Survival Analysis

This section compares the survival distribution generated by this machine learning algorithm to the survival distributions generated using the optimist and pessimist judgements. The predictions for the P(L|F,E) provided by the Use GBR full model algorithm were used to generate the survival distributions. The two types of validation, LOGO and LOO, were considered. The survival distributions presented in Figure 5 were obtained using Eq. (1). Figure 5 presents the survival distributions for the ice shelf environment for the *Autosub 3* AUV.

Visually, the survival plots show that there is a significant difference between the expert model and the GBR full optimist model for the optimist survival distribution for the ice shelf survival distribution. The difference between the survival distributions is smaller for the pessimist model. These observations are supported by X^2 significance tests. The null hypotheses is that there is no difference between the machine learning risk profile and the expert risk profile. For open water, USE-GBR-Full model, for the optimist prediction, LOO and LOGO obtained a $X^2 = 0.009$ and $X^2 = 0.0019$, respectively. Given that there is 38 degrees of freedom, the X^2 critical is 53. Therefore, we cannot reject the null hypothesis. The differences between the survival model generated by the expert and that generated by the USE-GBR-Full model are not statistically significant. The same conclusion can be reached

for the pessimist model for the open water. The LOO and LOGO obtained a $X^2 = 3.84$ and $X^2 = 3.11$, respectively.

For the ice shelf, USE-GBR-Full model, for the optimist, LOO and lOGO validation we have obtained a $X^2 = 0.130$ and $X^2 = 0.920$, respectively.

For the ice shelf environment, the results are better for the pessimist model when compared to the optimist model. The survival curves in Figure 5 show that the survival distribution of the USE-GBR-Full is very similar to the distribution obtained using experts' judgements. When we tested the differences between the experts' survival and the USE-GBR-Full survival distribution, we obtained the X^2 values of 0.0654 and 0.392 for the LOO and LOGO validations, respectively. These values are below the X^2 critical of 53. The differences between the survival distributions generated using the machine learning algorithms and the distributions generated based on expert judgements are not statistically significant.

Visually, from Figure 5, it is possible to see that for ice shelf environment the optimist performs worse than the pessimist. The explanation for this is that the variability in the aggregated probability estimations for the optimist model is higher than the variability for the aggregated pessimist model. For ice shelf, the aggregated optimist estimations for *Autosub 3* have a minimum value of 0.0005, a maximum value of 0.997, and a median of 0.005. The aggregated pessimist model, for ice shelf, has a minimum value of 0.104, a maximum value of 0.985, and a median of 0.542. A similar but reverse observation can be made for the open water environment which explains why, for this environment, the optimist model performs better than the pessimist model. For the open water environment the aggregated optimist estimations have a low variability when compared to the variability of the estimations of the aggregated pessimist model.

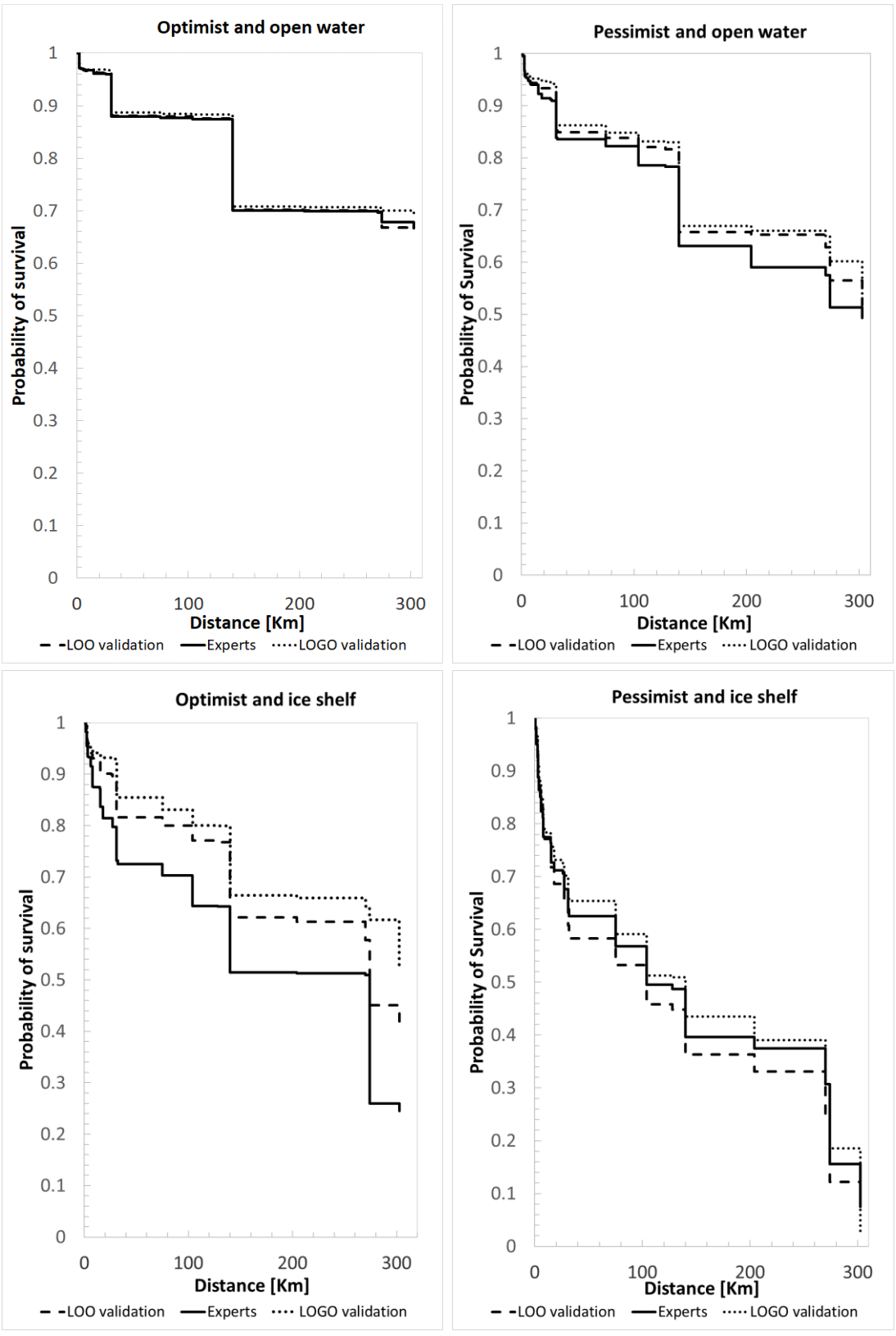

Figure 5: Subjective Kaplan Meier survival distribution for Experts, Use GBR Full model (LOGO validation) and Use GBR Full model (LOO validation). Top Left is Optimist and Open Water; Top ight is pessimist and open water. Bottom Left is optimist and ice shelf; Bottom Right is pessimist and ice shelf.

5 Limitations

To this date, probabilistic risk assessment relies on experts' judgements (O'Hagan et al., 2006). We have investigated whether machine learning algorithms can emulate the experts' judgements based on contextual and environmental evidence. In this paper, we have shown that this is possible. However, there are several limitations in our approach, and we discuss these in this section.

The accuracy of a machine learning algorithms presented in this paper is dependent on consistent data. Expert judgements are susceptible to biases such as the conjunction bias, base rate neglect, or others (Kynn, 2008). If not identified and corrected, biases can lead to inconsistencies in the assessments provided by the experts. Formal expert judgment elicitation goal is to mitigate potential biases. A limitation of our approach is that the machine learning algorithm must be applied to judgements that have been elicited using formal judgement elicitation methods. This involves a formal process for expert selection, expert training and expert judgements' aggregation (Keeney and Winterfeldt, 1991).

We have applied machine learning algorithms to reproduce the aggregated expert judgements. The expert judgements were aggregated considering optimist and pessimist models. This worked in the favour of the machine learning algorithms because it has reduced the variability of the training data. In practical terms, the variability in the estimation is a limitation in mathematical aggregation. Considering linear weighted aggregation, if there are two experts, one assigns a probability of 0.0001 and another assigns a probability of 0.1 to the same event. If the experts are allocated the same weight, the aggregated probability of the event occurring is 0.05. If, on the other hand, one expert assigns a probability of 0.05 and the second expert assigns a probability of 0.1, the aggregated probability of the same conditions is 0.075. When the experts' weights have the same weight, pessimist estimations can to some extent censor the optimist estimations. If one is considering weighted log aggregation, if one

expert assigns a probability of 0 to an event, then the aggregated probability is 0; in this case an optimist estimation censors the pessimist estimation. Variability in the estimations is a problem in the aggregation. In our view the proposed machine learning algorithms fitted the aggregated expert assessments well because these were aggregated in optimists and pessimists.

In this paper, we have not tested if the type of expert judgement aggregation affected the model's accuracy. We have used data from linear expert judgement aggregation and from behavioural expert judgement aggregation. Different types of expert judgement aggregation may introduce different problems. The linear and log pool aggregation is very sensitive to low- and high-probability judgements (Otway and von Winterfeldt, 1992) whilst behavioural expert judgement elicitation is susceptible to group polarisation(van Steen, 1992). Our data processes attempted to reduce potential errors by considering expert judgements where the weight provided by the expert was equal to or higher than 3. With respect to *Remus 100* elicitation data, we have used judgements provided for the 95 quantile. The judgements used in the model presented in this paper were of high confidence. However, there are some expert judgements' elicitations where there is significant uncertainty in the judgement. From our study we cannot confirm that machine learning algorithms are suitable for judgements where there is a significant amount of uncertainty.

Expert judgement reliability is an important problem in PRA and a source of much criticism (Bolger and Wright, 1994). Proposed solutions to this problem rely on the use of seed questions to assess the expert's reliability based on accuracy and information scores (Hanea and Nane, 2019). Seed questions are questions for which the facilitator knows the answer, but for which the experts do not know the answers but are asked to give estimates. Expert's performance with the seed questions can provide information about the expert's confidence and accuracy. The Excalibur expert judgement elicitation method uses this score

to calculate weights for each expert (Goossens et al., 2008). For the actual PRA the correct answer – the probability of an adverse event occurring – is unknown. Hence, there is an argument that seed questions are not an ideal indicator of how well an expert performs in the actual probabilistic risk estimation. The methodology presented in this paper does not address the problem of experts' reliability but it can address the problem of consistency and inform new weighting schemes for expert judgement aggregation. The machine learning algorithm is trained to give risk estimations for a given type of platform, in a given environment, for a given fault description. For a new fault, predictions made by the machine learning algorithms can be compared to predictions made by individual experts. The difference between the two predictions can be used to estimate a weight for the expert. Observation data from actual AUV missions can be used to test the reliability of both the expert prediction and of the machine learning algorithm.

The impact of risk mitigation on the risk profile can be captured using Bayesian models (Brito and Dawson, 2020). These models take as input the expert *a priori* estimate of the probability of a fault being mitigated. The probability of a fault being mitigated is not explored in this paper. Future research should explore whether or not it is possible to use narrative of the fault mitigation strategy to estimate the probability of a fault being mitigated by corrective action.

We argue that text mining and sentiment analysis can be applied to measure the consistency of the experts' assessments for a given problem.

6 Conclusion

Studies on expert judgement reliability have focused on comparing expert judgement estimates with actual observations. In our view, this presents two problems. First, it does not take into account the narrative associated with a given judgement. The narrative captures important information such as the context, which describes potential impacts and

environmental conditions. The second problem with the current approaches is that one can only validate expert judgements when data become available. When one needs expert judgement elicitation, the actual data are not available. This makes it impossible to assess the reliability of the expert judgements. In this paper we address these two problems.

When we attempted to model expert judgements using narrative information of the event description, we realised that we have developed a method for verifying the consistency of experts' judgements. We show that machine learning algorithms can identify individual words and groups of words that affect expert judgement with respect to risk estimation.

We have tested two machine learning algorithms and four text-embedding methods. The best performing algorithm, USE-GBR-Full, was then used to conduct a probabilistic risk assessment of a problem for which expert judgements are publicly available – the risk assessment for *Autosub 3* deployment in under the Pine Island Glacier in Antarctica.

When we compared the survival distributions obtained using the USE-GBR-Full machine learning algorithms, the results showed that the differences between the survival distributions of these algorithms and those generated by the experts are not statistically significant.

The machine learning algorithm was trained on the data of a *Autosub 3* and a *Remus 100*. These are two autonomous underwater vehicles in different classes. While *Autosub 3* is classed as a large AUV, over 3 metres in length and over 1000 Kgs in dry weight, *Remus 100* is a small AUV slightly over 1 metre in length and less than 40 Kgs in dry weight. Given that the machine learning algorithm was trained on small and large AUVs, the method can be used to develop *a priori* risk model for any other AUV given that a modicum of faults are collected and a full description of the fault and distance at which it has occurred is obtained.

Machine learning algorithms for probabilistic risk estimation can be applied to other domains; for example, health statistics. When combined with the subjective statistical survival estimator this method can quantify the risk of a catastrophic event as a function of

time. Narrative of the symptoms and health-related structural data (such as time under treatment, or other health complications) could be used to estimate the probability of a catastrophic event for a given individual under a given treatment. This would present a static risk model, trained on previous health experts' assessments, and the probability of a given catastrophic event would be computed for each individual. The subjective statistical survival estimator would take as input the probabilities for all health patients to compute the probability of survival over time for a population under study. The methodology presented in this manuscript can be useful for subject areas where the narrative may add value to the risk prediction power. The process is presented in this manuscript so others can easily tailor it to their applications.

References

Adam, M., Wessel, M. & Benlian, A. 2020. AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, 1-19.

AI Performance. 2020. Time for AI to cross the human performance range in chess. Accessed 2021-04-26. URL: https://aiimpacts.org/time-for-ai-to-cross-the-human-performance-range-in-chess/

Apostolakis, G. 1990. The Concept of Probability in Safety Assessments of Technological Systems. *Science,* 250**,** 1359-1364.

Aven, T. 2016. Risk assessment and risk management: Review of recent advances on their foundation. *European Journal of Operational Research,* 253**,** 1-13.

Bigün, E. S. 1995. Risk analysis of catastrophes using experts' judgements: An empirical study on risk analysis of major civil aircraft accidents in Europe. *European Journal of Operational Research,* 87**,** 599-612.

Bolger, F. & Wright, G. 1994. Assessing the quality of expert judgment: Issues and analysis. *Decision Support Systems,* 11**,** 1-24.

Bonano, E. J., Hora, S. B., Keeney, R. L. & Winterfeldt, D. V. 1990. Elicitation and Use of Expert Judgment in Performance Assessment for High-Level Radioctive Waste Repositories. U.S. Nuclear Regulatory Commission.

Brito, M., Griffiths, G. & Trembanis, A. 2008. Eliciting expert judgment on the probability of loss of an AUV operating in four environments. *Research and Consultancy Report.* National Oceanography Centre, Southampton.

Brito, M. P., Griffiths, G. & Challenor, P. 2010. Risk Analysis for Autonomous Underwater Vehicle Operations in Extreme Environments. *Risk Analysis,* 30**,** 1771-1788.

Brito, M. & Chang, Y. 2018. On the reliability of expert's assessments for autonomous underwater vehicle risk of loss prediction: are optimists better than pessimists?

Proceedings International Conference on Probabilistic Safety Assessment and Management (PSAM14), Los Angeles, California, pp 1-12.

Brito, M. & Dawson, I. 2020. Predicting the validity of expert judgments in assessing the impact of risk mitigation through failure prevention and correction. *Risk Analysis*, 40, 1928-1943

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C. & Sung, Y. H. 2018. Universal sentence encoder. Available: https://arxiv.org/abs/1803.11175.

Cooke, R. & Goossens, L. H. J. 2004. Expert judgment elicitation for risk assessments of critical infrastructures. *Journal of Risk Research* 7**,** 643-656.

De Campos, L. M. & Castellano, J. G. 2007. Bayesian network learning algorithms using structural restrictions. *International Journal of Approximate Reasoning*, 45**,** 233-254.

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. . Available: https://arxiv.org/abs/1810.04805.

Domingos P., & Pazzani, M. 1996. Beyond independence: conditions for the optimality of the simple Bayesian classifier, Presented at the 13th International Conference on Machine Learning, Bari, Italy.

Fan, W., Wallace, L., Rich, S. & Zhang, Z. 2006. Tapping the power of text mining. *Communications of the ACM* 49**,** 76–82.

Ferrari, S. & Cribari-Neto, F. 2004. Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31**,** 799-815.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29**,** 1189-1232.

Garcia, D. 2013. Sentiment during recessions. *The Journal of Finance*, 68**,** 1267–1300.

Garthwaite, P. H., Kadane, J. B. & O'Hagan, A. 2005. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association,* 100**,** 680-701.

Goldberg, Y. & Levy, O. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method arXiv preprint arXiv:1402.3722. Available: https://arxiv.org/pdf/1402.3722.pdf.

Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. 2016. *Deep learning (Vol. 1),* Cambridge, Massachusetts, MIT Press.

Goodwin, P., Gönül, M. S. & Önkal, D. 2019. When providing optimistic and pessimistic scenarios can be detrimental to judgmental demand forecasts and production decisions. *European Journal of Operational Research,* 273**,** 992–1004.

Goossens, L. H. J., Cooke, R. M., Hale, A. R., & Rodic-Wiersma, L. 2008. Fifteen years of expert judgement at TUDelft. *Safety Science,* 46(2), 234–244.

Griffiths, G., Brito, M., Robbins, I. & Moline, M. 2009. Reliability of two REMUS-100 AUVs based on fault log analysis and elicited expert judgment. In Proceedings of the International Symposium on Unmanned Untethered Submersible Technology. *UUST 2009 Durham NH, USA,*. Durham, New Hampshire Autonomous Undersea Systems Institute.

Groth, S. S. & Muntermann, J. 2011. An intraday market risk management ap- proach based on textual analysis. *Decision Support Systems,* 50**,** 680–691.

Hanea, A. M. & Nane, G. F. 2019. Calibrating experts' probabilistic assessments for improved probabilistic predictions. *Safety Science,* 118**,** 763-771.

Hänninen, M. & Kujala, P. 2012. Influences of variables on ship collision probability in a Bayesian belief network model. *Reliability Engineering & System Safety,* 102**,** 27-40.

Head, T., Mechcoder, G. L. & Shcherbatyi, I. 2018. *Scikit-optimize*: v0.5.2.

Hodge, V. J. & Austin, J. 2004. A Survey of Outlier Detection Methodoligies. *Artificial Intelligence Review,* 22**,** 85–126.

Holton, C. 2009. Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decision Support Systems,* 46**,** 853–864.

Kabir, G., Tesfamariam, S., Francisque, A. & Sadiq, R. 2015. Evaluating risk of water mains failure using a Bayesian belief network model. *European Journal of Operational Research,* 240**,** 220-234.

Kahneman, D. & Tversky, A. 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychology,* 3**,** 25.

Kaplan, E. L. & Meier, P. 1958. Nonparametric estimation from incomplete observations. *J. American Statistical Associations,* 53**,** 457-481.

Kaplan, S. & Garrick, B. J. 1981. On the Quantitative Definition of Risk. *Risk Analysis* 1**,** 11-27.

Keeney, R. L. & Winterfeldt, D. V. 1991. Eliciting Probabilities from Experts in Complex Technical Problems. *IEEE Transactions on Engineering Management,* 38**,** 191-201.

Kynn, M. 2008. The 'heuristics and biases' bias in expert elicitation. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 171**,** 239-264.

Landauer, T. K., Foltz, P. W. & Laham, D. 1998. An introduction to latent semantic analysis. *Discourse Processes,* 25**,** 259-284.

Lee, J. & Yi, J.-S. 2017. Predicting Project's Uncertainty Risk in the Bidding Process by Integrating Unstructured Text Data and Structured Numerical Data Using Text Mining. *Applied Science,* 7**,** 1-15.

Lessmann, S., Baesens, B., Seow, H.-V. & Thomas, L. C. 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research,* 247**,** 124-136.

Liu, Y.-H. & Chen, Y.-L. 2018. A two-phase sentiment analysis approach for judgement prediction. *Journal of Information Science,* 44**,** 594-607.

Medvedeva, M., Vols, M. & Wieling, M. 2019. Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*. 28, 237–266.

Merkhofer, M. W. 1987. Quantifying judgmental uncertainty: Methodology, experiences, and insights. *IEEE Transactions on Systems, Man, and Cybernetics,* 17**,** 741-752.

Mohammad, S. M. & Turney, P. D. 2010. Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text.

Morris, A. P. 1977. Combining expert judgements: A Bayesian approach. *Management Science,* 22**,** 679-693.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. & Rakow, T. 2006. *Uncertain judgements: Eliciting experts' probabilities*, Chichester, UK, Wiley.

Otway, H. & Von Winterfeldt, D. 1992. Expert judgment in risk analysis and management: process, context, and pitfalls. *Risk Analysis,* 12**,** 83-93.

Paltrinieri, N., Comfort, L. & Reniers, G. 2019. Learning about risk: Machine learning for risk assessment. *Safety Science,* 118**,** 475-486.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. 2018. Deep contextualized word representations. Available: https://arxiv.org/abs/1802.05365.

Ribeiro, R. T., Singh, S. & Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144.

Sarkar, S., Vinay, S. & Maiti, J. 2016. Text mining based safety risk assessment and prediction of occupational accidents in a steel plant. *International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT).* New Delhi.

Sigurdsson, J. H., Walls, L. A. & Quigley, J. L. 2001. Bayesian belief nets for managing expert judgment and modelling reliability. *Quality and Reliability Engineering International,* 17, 181-190.

Son, B.-Y. & Lee, E.-B. 2019. Using Text Mining to Estimate Schedule Delay Risk of 13 Offshore Oil and Gas EPC Case Studies During the Bidding Process. *Energies,* 12**,** 1956, 25 pp.

Strut, J. 2006. *Report of the inquiry into the loss of Autosub2 under the Fimbulisen*. National Oceanography Centre, Southampton.

Tversky, A. & Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science,* 185**,** 1124-1131 .

Van Steen, J. F. J. 1992. A perspective on structured expert judgment. *Journal of Hazardous Materials,* 29**,** 365-385.

Vapnik, V. 1995. *The nature of statistical learning theory*, New York, Springer.

Yadav, R., Kumar, A. & A, V. K. 2019. Event-based sentiment analysis on futures trading. *Journal of Prediction Markets,* 13**,** 57-81.

Winkler, R.L., 1968. The Consensus of Subjective Probability Distributions. *Management Science,* 15(2), B61-B75.

Zou, H. & Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 67**,** 301-320.