Mixed Initiative Approach for Reliable Tagging of Maintenance Records with Machine Learning

Naresh Iyer[1], Nurali Virani[1], Zhaoyuan Yang[1], and Abhinav Saxena[1]

[1] *GE Research, Niskayuna, NY, 12309, USA*
iyerna@ge.com
nurali.virani@ge.com
zhaoyuan.yang@ge.com
asaxena@ge.com

ABSTRACT

Free-form text-based maintenance and service records related to industrial assets capture the observations and actions of service engineers and are a crucial resource for assessing system-level asset health. To facilitate tracking of historical asset health issues, these records are categorized using tags from a predefined taxonomy, which is mostly a manual and time-consuming process. Given that these records can offer valuable information in troubleshooting maintenance issues, automating this process through deep learning (DL) based natural language processing (NLP) models can offer significant operational and maintenance (O&M) cost reductions. However, these data-based models are not expected to be fully accurate, requiring human experts to regularly review all predictions by DL models to verify or correct them, which is also an highly inefficient and costly process. On the other hand, new records that have novel or ambiguous context can be more appropriately resolved by a human expert. The objective of the work described in this paper is to create an interpretable mechanism that can assess reliability of individual predictions from DL-based maintenance record classifiers and help design a mixed initiative system. This system aims to identify scenarios where predictions are reliable enough for automated decision versus where human intervention is needed due to poor reliability. Additionally, this system aides decision support by providing exemplars from training set that can enhance the human tagger's productivity and quality. Given a set of tagged records, it also has the capability to identify instances where the originally assigned tags are likely to be inaccurate/noisy. We illustrate these outcomes through tagging of maintenance records from the aviation domain, leading to improvements over only human-based or only DL-based tag assignments.

1. INTRODUCTION

With the advent of digital tools, industrial systems are increasingly exploring ways to further improve industry KPIs such as reduction in operational and maintenance (O&M) costs, efficient and effective field services, speed of resolution and minimized downtimes. Methods to access text-based knowledge have become the topic of mainstream research within the PHM community. Digitized knowledge in today's world is often required to be codified in semi-structured form, that is not fully tabular in nature – some examples include full-text repositories, maintenance and service records, call center records, internet webpages, Powerpoint documents, service manuals, and expert annotations. Many such repositories, however, are still organized to be queryable using abstracted meta-fields or tags which are supposed to loosely organize the overall information content into appropriate categories. Tags allow for efficient browsing, querying and extraction of relevant records from the repositories. In this paper, we look at an instance of this problem as applicable to maintenance and service records related to industrial assets. These records capture the observations and actions of service engineers and are a crucial resource for assessing system-level health of an asset and for inferring reliability issues arising from those. They contain useful information such as a description of the issue addressed and references to the actions and observations generated during resolution of the service event. They capture the decision-making and actions of service engineers and are an extremely critical resource for our service engineers to help make quick and accurate decisions. Typically, these records are semi-structured and free text is used to describe observed issues and relevant corrective actions that were performed in response. To facilitate tracking of historical issues related to various system components, these records are categorized using tags from a predefined domain taxonomy. For example, in the Aviation industry, one of such tagging taxonomy is based on the Air Transport Association of America (ATA) developed coding

systems that serve as a common standard for sharing information between various technical personnel related to commercial aircraft: the *ATA iSpec 2200* is an industry-wide for aircraft system numbering (ATA Chapters), whose main objectives are stated to be "to minimize cost and effort expended by operators and manufacturers, improve information quality and timeliness, and facilitate manufacturers' delivery of data that meet airline operational needs". For example, chapter *ATA 32* pertains to records that relate to the Landing Gear of the aircraft, while chapter *ATA 73* is used to tag records related to Engine Fuel and Control. These tags or ATA Chapters enable efficient retrieval of information most relevant to a current scenario, which can then inform the optimal response to the scenario at hand, based on historical resolution of similar scenarios. As yet another example, the Nuclear industry is focused on reducing O&M costs (GEMINA, 2019), due to significant manual work that is performed in assessing plant condition and generating work orders. A similar tagging framework is expressed as classification of systems into Functional Equipment Groups (FEG). FEGs are assigned based on maintenance rules, applicable equipment tag lists, drawings, procedures, previous tag outs, etc. within an organization, FEGs are assigned and reassigned to optimize maintenance efficiency as processes evolve over time. Experts depend on *tribal knowledge*; moreover, due to the absence of an industry-wide standard like ATA Chapters, experts in the Nuclear industry have to often sift through extensive documentation in distilling a succinct and efficient set of steps to be executed. Therefore, optimizing a taxonomy for existing maintenance processes and further using an existing taxonomy for tagging of text-based information are both critical research problems towards transforming another heavily regulated industry, like Nuclear, that is highly relevant towards producing decarbonized energy(GEMINA, 2020). In this paper we focus on the reliable automated tag assignment problem, using the Aviation industry as example, where an underlying taxonomy such as ATA chapters exists and publicly available.

The benefits of tagging unstructured data can only follow if tags are assigned accurately and reliably each time there is a new record to be categorized. In many applications, human experts are employed to assign and curate tags, making it a highly inefficient, costly and error-prone process. The logical approach for tackling this issue involves replacing the manual expert with an AI-based automated inference engine that can learn patterns from the existing tagged data to automatically assign tags to a new record. New advances in Natural Language Processing (NLP) using Deep Learning (DL) (Min et al., 2021) have shown remarkable capabilities in the analysis of unstructured text data, allowing for automation of complex activities like question answering, next sentence prediction, and document summarization. Using the existing repository of unstructured service records and tags that were assigned to them, a supervised learning approach can be implemented to automate the tagging process, using Deep Learning based language models like BERT (Devlin, Chang, Lee, & Toutanova, 2019). However, while the overall statistical accuracy of such models have been shown to be high, it is not often clear how to assess whether an individual prediction from such a model can be trusted. For instance, the tags based on predictions produced by such a model when the input records that have newer or ambiguous context can be unreliable, or lack enough confidence to result into actionable decisions.

On one hand, we expect DL models to provide reliable predictions for samples that are similar to the training samples used to train the model. At the same time, humans are inclined to be relatively superior at tackling exceptional cases. For e.g., new records that have novel or ambiguous context are better candidates for resolution by a human expert, compared to a DL model. Based on this insight, we target the design of a mixed initiative tagging system that can assess reliability of individual predictions of machine learning models - a goal of the Humble AI initiative at GE. Specifically, we make use of a *Justification-based* reliability assessment of individual predictions by DL classifiers (Virani, Iyer, & Yang, 2020), which we call Epistemic Classification. Using this approach, we ascribe reliability to tag predictions made by our BERT-based, automated tagging model; the reliability assessments further help to identify cases to assign to human experts for tagging, versus those that can be reliably tagged automatically. In other words, the mixed initiative system aims to identify scenarios where automated tag predictions can be trusted in contrast to ones where human intervention is warranted due to unreliability of model prediction. This results in the optimal division of labor and trade-off between employing expensive manual labor and obtaining accurate tag predictions. The system also generates information that contributes to the interpretability of its outcome, which is valuable for developing trust in its predictions. The interpretability-relevant information is also used to cue the human decision maker optimally in cases where a human is chosen to perform the tagging; this helps in enhancing the productivity and quality of the human tagger's effort. Although one might try to determine prediction reliability by choosing a threshold on softmax values from the final layer of a DL model, these values rely only on distance from the classification hyperplane and do not consider impact of extrapolation or overlapping training distributions. Moreover, softmax thresholding approach does not provide any interpretability. An additional benefit of our system is its ability, given a set of tagged records, to identify instances where the originally assigned tags are likely to be inaccurate. Finally, we illustrate outcomes from applying our approach on public text data sources and demonstrate reliable tagging of maintenance records from Aviation domain, leading to improvements over a purely human-based or a DL-based tagging approach.

2. PROPOSED APPROACH

Figure 1 shows the overall workflow based on our approach. For the baseline model, historical semi-structured records, along with tags that were assigned to them, are provided as training data to a BERT-based classification model for supervised learning of tags. We consider two variations for baseline: baseline model 1 (BM1) directly uses the pretrained BERT-based model for inference, using the feature embedding to perform tag prediction using dense classification layers. To further improve classification performance, a standard practice is to fine-tune the model, including the inference layers, to the current problem (dataset) - this is baseline model 2 (BM2). Each of these 2 baseline models are then evaluated using our Epistemic framework, that helps assess reliability of individual tag predictions, thereby identifying (as show in Figure 1) three categories of assessments: High, Medium and Low confidence predictions. These reliability categories are further used to estimate when the automated tag assignment requires intervention by a human expert. We show multiple mechanisms by which the Epistemic framework enhances the overall quality of the decision making involved in accurate and reliable assignment of tasks to the service records. Specifically, we show that in more than 50% of the opportunities (for the dataset we used in our experiments), the tagging can be fully automated while preserving the accuracy of the assignment with high reliability. Conversely, our approach also identifies service records where the potential for a model, even with high statistical accuracy, to assign an incorrect tag is high, thereby warranting the need for intervention by a human expert to resolve the case and assign the right tag. Additionally, the system provides evidence justifying why these service records are difficult from the perspective of the model. We describe details of the mixed initiative system next.

2.1. Problem description and challenges

Aircraft maintenance and service records are critical for maintaining airworthiness of an aircraft; they carry details related to the repair performed on the aircraft. These records capture the observations and actions of service engineers and are a crucial resource for assessing system-level health of an asset and for inferring reliability issues arising from those. Typically, in these records, free text is used to describe observed issues and relevant corrective actions that were performed in response - Figure 2 shows snapshots of 2 such examples of service records as captured in (FAA, 2001). As can be seen, in each example, a free text entry describes critical elements of issues discovered as well as symptoms observed, and upon repair the corrective action (or set of actions) that was performed to fix the issue. A vital element of the lifecycle of this record involves assigning it a specific tag (or ATA Chapter) from a predefined taxonomy of ATA chapters. In today's world, a large volume of these records is created daily; for instance, it is estimated that over 1000 such records are created daily for a fleet of 300 aircraft. Due to their volume, the manual processing of these free text records to assign an accurate tag, can be a laborious task, given the number of ATA chapters are in the 100s and each chapter can have 10's of sub-chapters. In addition to the cost, manual processing is also error-prone and can likely lead to incorrect assignment of tags, thereby making the repository less effective in terms of enabling accurate browsing, querying and knowledge reuse capabilities. Therefore, automating the assignment of tags from the content of service records is a valuable capability and a necessity in today's industry.

2.2. Inference Architecture

Recent advances in NLP and the use of deep learning models for processing unstructured documents and text has shown remarkable advances and are well-suited for the task of automated tag assignment. As a first step of our approach, we develop and show outcomes from the application of one such model for the tag assignment task. Typically, an unsupervised language model is trained using a large corpus of data and then fine-tuned on the downstream task. Multiple instances of such language models exist in literature including ELMo (Embeddings from Language Models) (Peters et al., 2018), ULMFiT (Universal Language Model with Fine-tuning) (Howard & Ruder, 2018), OpenAI GPT (Generative Pre-Training) (Radford & Narasimhan, 2018), BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) and OpenAI GPT-2 (Radford et al., 2019). We leverage the BERT pre-trained model as made available by Google (Devlin, Chang, Lee, & Toutanova, 2018). BERT makes use of an encoder-decoder architecture that contains Transformers (Vaswani et al., 2017), which utilizes an attention mechanism to extract features for each word in a sentence in a way that leverages the order, sequence as well as position of all other words in the same sentence, i.e., it learns language-specific context. Once the model is pre-trained on a given language, the embedding within the encoder-decoder architecture of the pretrained model can be directly used to attend to multiple tasks within the language, often without requiring further training for the task at hand. For our task, we utilized the BERT-Tiny architecture to process the tokenized maintenance and service records, learning to classify tags by making use of the dense classification layers that we connect to the BERT embedding. This is shown in Figure 3. While BERT-Large and BERT-Base are more commonly used as pretrained models in the community, we employed BERT-tiny that is also made available by Google for experimentation with language models on a smaller scale; in the current application, we intended to fine-tune the pretrained model parameters using the service records, as one of the 2 baselines. Due to the heavy computational footprint of fine-tuning the larger 2 versions of BERT (see Figure 4), we decided to perform our experiments with BERT-tiny, ap-

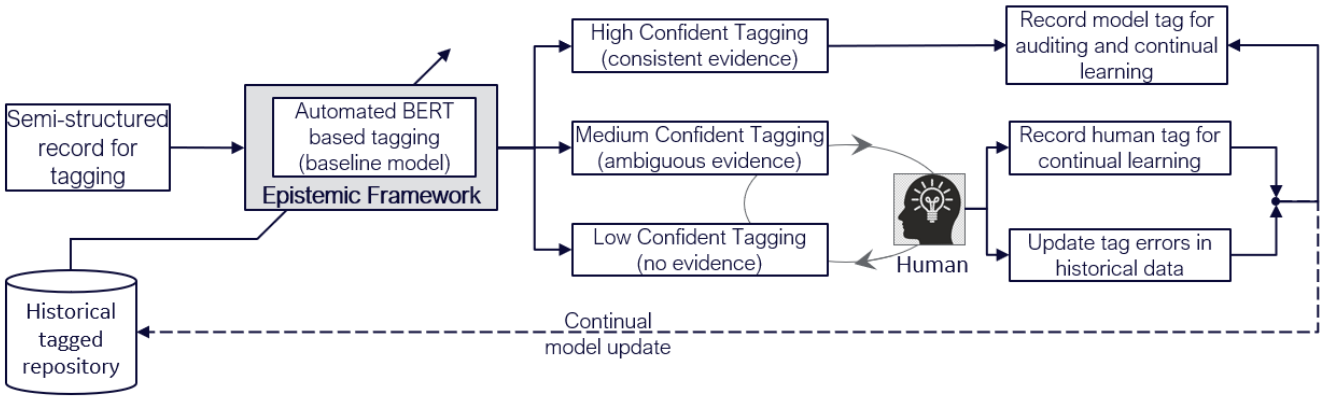

Figure 1. The overall mixed initiative approach for reliable tagging of maintenance service records.

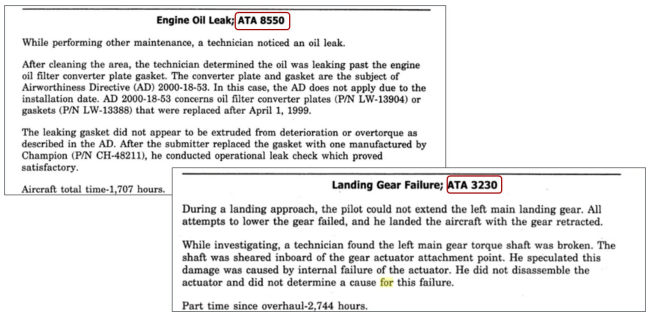

Figure 2. Two examples of maintenance/service records with red boxes indicating the ATA Chapters (tag) assigned to each.

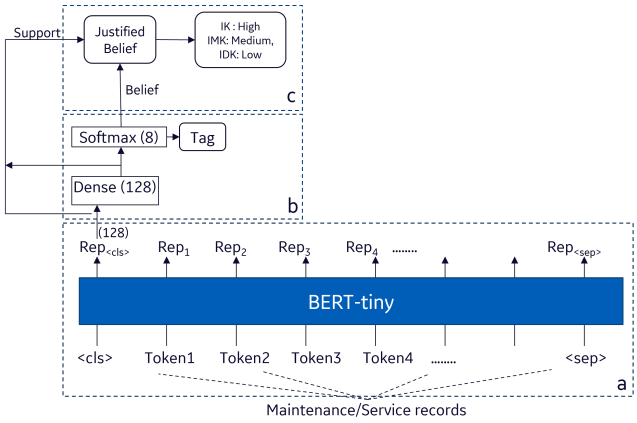

Figure 3. The BERT-tiny architecture with elements of the epistemic framework to generate reliability assessments for tag predictions. Boxes a, b and c indicate the primary elements of the architecture - a is the pretrained BERT model, b includes additions to perform classifications from the embedding of the pretrained model, and c shows extensions that make use of the overall architecture to generate epistemic reliability assessments for individual tag predictions by the model.

plying 128 tokens, which has far fewer parameters than its 2 larger counterparts. For the classification layers, we used a dense layer with 128 neurons, followed by a 'softmax' layer for the multi-tag classification. We next describe details related to the service records data that we used for our experiments.

2.3. Data description

As mentioned previously, the data comprises of unstructured text records that contain comments from service, field and maintenance engineers. By way of the current case management system, maintenance staff is also required to assign a tag to the record based on problem description, which is both time consuming and error prone and is consequently targeted by this work. However, for initial system development the ATA Chapter tag assigned to each record is leveraged (as illustrated in Figure 2), which is considered the ground truth for the tag assignment (classification) problem. Our dataset had a total of 142,936 records from which we retain only those samples for which the assigned tag occurs at least twice in the dataset; this brings the number of samples in the dataset to 142,813. Additional pre-processing was required to eliminate commas and non-alphanumeric characters in the text. This results in a dataset with 454 unique ATA Chapter tags, making it a 454-class problem. Figure 5 shows the distribution the tags for the entire dataset. We find that 8 tags alone cover about 40% of the overall dataset (or, about 55,000 samples); as a result we reformulate the original problem as an 8-class problem, targeting only the most frequent tags occurring in the data. The imbalance in the 8 classes is shown in terms of the percentages in the second column of the table. We would like to point out that this filtering to the top 8 tags was done to make the underlying classification problem less complex, since the primary goal of this paper is to demonstrate a system that can contribute to reliability of the classification for individual predictions, in contrast to building the optimal classifier. That said, we believe the concepts and the system demonstrated in this paper can be extended, without loss of generality, to the original 454-class

BERT Config	# encoders (L)	#embedding (H)
BERT-tiny	**2**	**128**
BERT-mini	4	256
BERT-small	4	512
BERT-medium	8	512
BERT-base	12	768
BERT-large	24	1024

Figure 4. Various configurations of the BERT architecture available for use; for reduced computational footprint we used BERT-tiny.

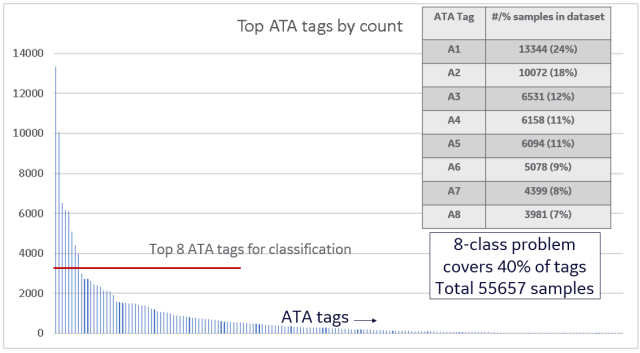

Figure 5. Showing the overall distribution of ATA tags in our dataset; we filter to the top 8 tags, for our experiment, that covers 40% of the dataset.

problem, although, the larger, imbalanced class dataset will adversely impact class separability, thereby challenging the baseline classifier as well as the fraction of samples that can be tagged with high reliability, and thus automated, by applying our Epistemic framework. The impact of large-class classification problems on the effectiveness of our Epistemic framework is a direction that is being pursued as future work. Before we describe details of our experiments and outcomes, we provide a brief overview on the idea of Epistemic Classification and its utility for assessing reliable classifications, or classification reliability.

2.4. Epistemic Classification

Epistemic classification is an approach, within the Humble AI initiative, inspired from the theory of Justified True Belief (JTB) in Epistemology ((Steup, 2007) is a good exposition), which aimed to study the limits and validity of human-acquired knowledge. We extend the same concept to understanding and characterizing the validity and limits of knowledge as acquired by supervised classifiers, as detailed in (Virani et al., 2020). We showed that the JTB analysis can be leveraged to expose the uncertainty of a classification

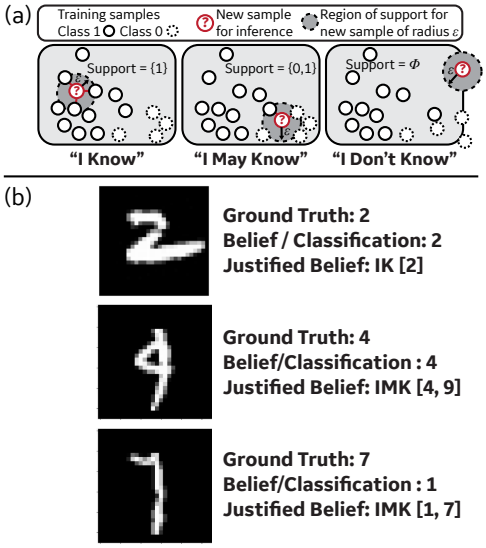

Figure 6. (a) Overview of justified belief using support from local neighborhood, (b) Illustration of the concept through a popular hand digit recognition problem.

model with respect to its inference due to ambiguity or extrapolation, thereby allowing for the inference to be only as strong as the justification permits. Through experiments conducted on simulated and real datasets, we demonstrated that our approach can provide reliability for individual predictions and characterize regions where such reliability cannot be ascertained. While specific details of the approach are in the paper (Virani et al., 2020), Figure 6 shows the primary idea of gathering epistemic evidence, along with some examples of applying epistemic classification to the digital classification problem using the MNIST dataset (Deng, 2012). The epistemic status of the classification for an individual prediction (within {I-Know, I-May-Know or I-Dont-Know}) is inferred based on the nature of support, from the training data, that the sample garners in its neighborhood, within one or more latent spaces of the neural-network based classifier. If the support shows uniform support, from the neighboring training samples, for the same class as the one predicted by the classifier, then the prediction is deemed as reliable and assigned an epistemic status of "I-Know". Alternatively, if such support has confusion between two or more classes, where one of those classes is the prediction, then the framework expresses doubt and even if its prediction is a single class, the epistemic status includes a union of all classes found in the support (as illustrated by the final 2 examples of Figure 6(b)).

We apply the same principle to the ATA tag classification problem to assess the reliability of a tag assignment to a service record. Through our experiments we show that tags that get assigned an epistemic status of "I-Know" tend to be highly reliable classifications - this has been seen and shown to be true based on the high classification accuracy of all the "I-Know" tag assignments; we will show the same to be the

case for the ATA tag classification problem. This outcome enables the natural design of a mixed-initiative system within which the "I-Know' tag assignments are treated as automated tag assignments without a need for supervision by a human expert; conversely, the less reliable tag assignments from the classifier are routed to the human expert for tag classification. The epistemic classification framework has a continuous parameter, ϵ, which defines the extent of the neighborhood within which support is generated for a given sample. This parameter serves as a trade-off parameter between two quantities: the proportion that defines how labor is divided between full automation and human intervention, and the desired strength of reliability required for a fully automated tag assignment. As one would expect, the more reliable that we want the fully automated tag assignment to be, the more samples will need to be rerouted to the human expert for intervention. Since the fraction of tag assignments that get routed via full automation are a good measure of efficiency of the mixed initiative system, we track this number as a metric and we call it *coverage*. We next describe the experiments and the outcomes based on this mixed initiative system for tag assignment.

3. EXPERIMENTAL SETUP

For our experiments, we invoke the BERT-based architecture shown in Figure 3 in multiple ways to serve as the baseline cases as well as instantiations of epistemic classification. As mentioned previously, we consider two variations of the architecture for use as baseline: baseline model 1 (BM1) directly uses the pretrained BERT-based model for inference, using the feature embedding to perform tag prediction utilizing dense classification layers. To further improve classification performance, a standard practice is to fine-tune the entire model to the current problem (dataset) - this is baseline model 2 (BM2). Each of these 2 baseline models are then evaluated using our Epistemic framework, which we call *Epistemic BERT*. Finally, the Epistemic classifier itself is additionally evaluated at multiple settings of parameter ϵ, which defines the size of the neighborhood within which support is sought for a given sample being evaluated. While multiple options exist for defining support, namely k nearest neighbors, ϵ radius, k nearest neighbors within ϵ radius, in the current work we make use of k nearest neighbors as the support operator: for each sample, *s*, being evaluated for tag assignment, we identify the k nearest training samples to *s*, in the identified latent space of the classifier, and assess the epistemic status of the tag assignment to *s* as a function of the ATA tags assigned to those neighboring training samples, using the inference framework indicated in Figure 6(a). In each case, we split the data (55k+ samples across eight classes shown in Figure 5) using an 80-20% split for training and validation respectively. Manually generated class labels originally provided with historical data were used to perform a stratified split.

3.1. BM1: Baseline BERT without fine-tuning

For BM1, the service records are first tokenized, using the tokenizer provided with the BERT distribution, and fed to the encoder of the pre-trained BERT model. This encoding is directly used, along with the tag assignments for the records, as input to a shallow classifier (like classification layers shown in Box b of Figure 3) to generate a model for tag assignment . The model is trained for 30 epochs and used to process new service records and directly predict ATA tags to assign them.

3.2. BM2: BERT with fine-tuning

It is generally accepted that the early layers in BERT's language model capture generic linguistic patterns that likely has little relevance to the downstream task (i.e., ATA tag assignment), while the later layers learn task-specific patterns. This intuition is derived from the same effect seen in deep computer vision models, where the initial layers learn generic features like edges and corners, while the later layers learn specific features, that are critical to a downstream task such as facial recognition. In line with this intuition we trained an alternate baseline model, by fine tuning the pretrained BERT model, using the service records data. More specifically, we run 80 additional epochs of training on all parameters of the pretrained model using our dataset of service records. Our results show that fine tuning has a significant impact on the ability of the model to learn patterns critical to accurate tag assignment.

3.3. Epistemic BERT

For each of the 2 baseline models (BM1 and BM2), we study 2 corresponding models that are epistemic classifier versions of the respective models. More specifically, for each tag prediction made by a baseline model we make use of *support* generated from two embeddings that are present in the classifier architecture - these embeddings include the feature embedding produced by BERT, fed as input to the Dense classification layer, and the output of the Dense layer for softmax-based class inference (this is illustrated in Figure 3 by the 2 arrows going into the block labeled *Justified Belief*). We make use of k-NN (k-Nearest Neighbors) as the support operator; in other words, a representation of a test sample in each of the 2 embeddings is created and 'k' nearest neighbors within the training data are retrieved. The tags assigned to the retrieved neighbors are then used, in conjunction with tag prediction made by the model for the test sample, to ascertain the epistemic status of the prediction (more details of this approach are described in (Virani et al., 2020)).

4. RESULTS

We present outcomes for the 4 models using a Class Confusion Matrix (CM) for the 8-class problem, that is typically used to indicate prediction performance of a classifier. For the epistemic models, (Virani et al., 2020) introduced an Augmented Confusion Matrix, a variant of CM, which splits it into 3 submatrices, each of which is a confusion matrix, but separately deals with predictions that have been assigned different epistemic statuses. In other words, the *I-Know* predictions are grouped within a dedicated confusion matrix showing performance of the classifier for the cases where the model's epistemic uncertainty is minimal and the predictions are highly confident. Looking at this submatrix gives us an insight into two metrics:

1. Accuracy (IK): The accuracy of the classifier for tag assignments that are designated as *I-Know* leading to automated assignment, and
2. Coverage (IK): The coverage of the classifier is the fraction of records that are designated as *I-Know* leading to automated assignment and thus, it represents the efficiency of the mixed initiative workflow.

Figure 7 compares the outcomes for the models BM1 and $BM1_\epsilon$, namely BERT-base and Epistemic BERT where there is no fine-tuning involved. The 8×8 confusion matrix on the left shows the performance of the base, pretrained BERT-based classifier on the service records, where the training data is used to tune only the classification layers (Box 'b' in Figure 3). The overall statistical accuracy of the model is 80% applied to the entire test dataset (i.e., for a coverage of 100% or full automation of the tagging process using the classifier). The 80% figure is at a performance level that is inadequate for an automated deployment in regulated industries, such as Aviation and Nuclear, since this would lead to 20% of the cases, on average, being tagged incorrectly.

The expected impact of applying the epistemic framework to drive a mixed initiative workflow for tagging is seen by looking at the augmented confusion matrix on the right of the figure. The confusion submatrix enclosed by the red box in Figure 7 contains the fraction of samples for which the epistemic uncertainty, by application of the model, is seen to be the lowest (*I-Know*). This fraction is 51% of the overall dataset. In other words, this result indicates that 51% of the records that we will see in the future can be expected to be automatically tagged because their epistemic status is *I-Know*. The remaining 49% of the records would need to be channeled to a human tagging expert. Moreover, upon examination of the confusion submatrix, we also see that the expected classification accuracy for predictions in this submatrix is 97%. While the overall system performance based on labeling and feedback from the human expert in our mixed initiative system was outside the scope of this paper, if we were to assume the human expert is 90% accurate, we get a much enhanced classifier with an accuracy of almost 94%, with more than 50% of the cases that can be reliably automated without human intervention. Yet another compelling argument is related to the significant reduction in human workload - only 49% of the records now require manual tagging as opposed to the baseline system where 100% of records are tagged manually.These benefits clearly indicate the value of Humble AI's Epistemic Classification approach when applied to the base classifier.

Figure 8 compares the outcomes for the models BM2 and BM2_ϵ, namely BERT-base and Epistemic BERT where the entire pretrained BERT model is further fine-tuned using the training data comprising of service records. In this case, we see a significant improvement in the performance of the base-BERT classifier, with an accuracy of 94%. While this statistical performance can be more compelling case to use the model in a fully automated workflow for tagging, the augmented confusion matrix provides an additional dimension that can motivate a mixed initiative system as a better alternative. Using a 3-NN support operator, the *I-Know* submatrix (enclosed by red box) of the augmented confusion matrix shows that 83% of the records can be tagged automatically based on their epistemic status being *I-Know* at an accuracy as high as 98%, with the remaining 17% of records that would be channeled to a human tagging expert.

Figure 9 shows outcomes for the tag classification problem under varying ablation conditions. Of particular interest are the final 2 columns of the table, which indicate the trade off between *Coverage* and *Accuracy* for the different versions of the epistemic classifiers. The table also shows how varying the size of neighborhood for estimation of *support* (i.e., the number of nearest neighbors, 'k' in the latent spaces considered for this set of experiments) helps exercise such a trade off. Changing 'k' from 3 to 10 results in a more stringent condition of justification, leading to a reduction in *coverage*, while at the same time increasing the accuracy of the *I-Know* matrix that will be the subset of cases that would be tagged in a fully automated manner. Based on the application, the end-user can specify their appetite for incorrect classification by the automated classifier by choosing an appropriate value of 'k' - while a more conservative (larger) value of 'k' implies better accuracy for the fully automated cases, it will also lead to a larger fraction of the samples being routed through a human tagging expert. Another interesting observation is the change in the magnitude of the *coverage* as a function of how good the base BERT model is, in terms of its accuracy. It seems to imply that a superior baseline model will provide superior levels of trade off between performance and coverage for the problem.

Figure 10 shows a comparison of the two baseline classifiers (with and without fine-tuning) with the corresponding 2 baseline Epistemic classifiers, in terms of class-specific precision,

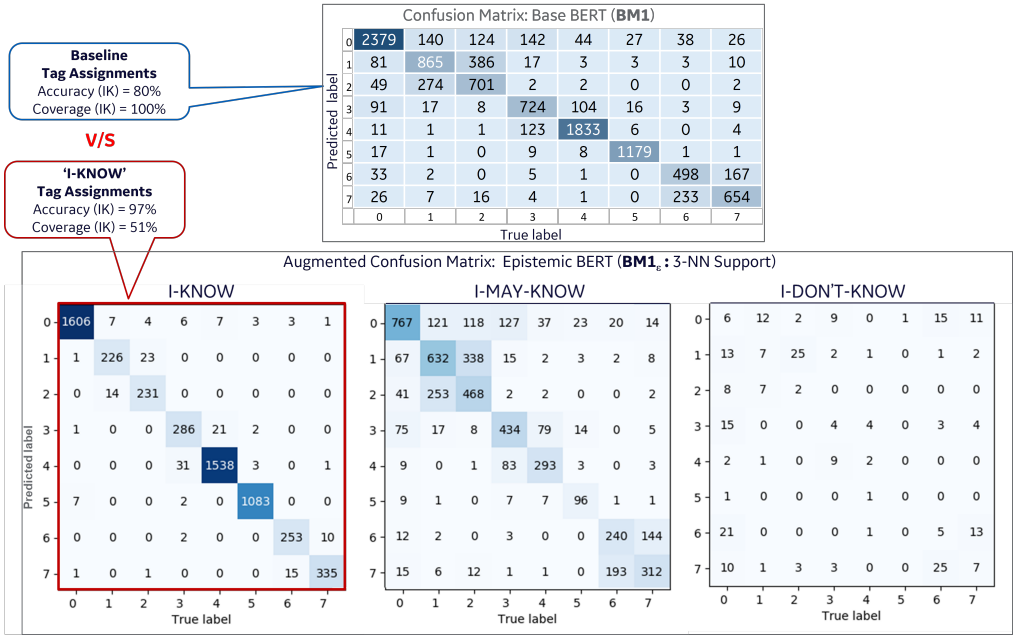

Figure 7. Comparing performance of the classifier for the case without any fine tuning applied to the pre-trained BERT model.

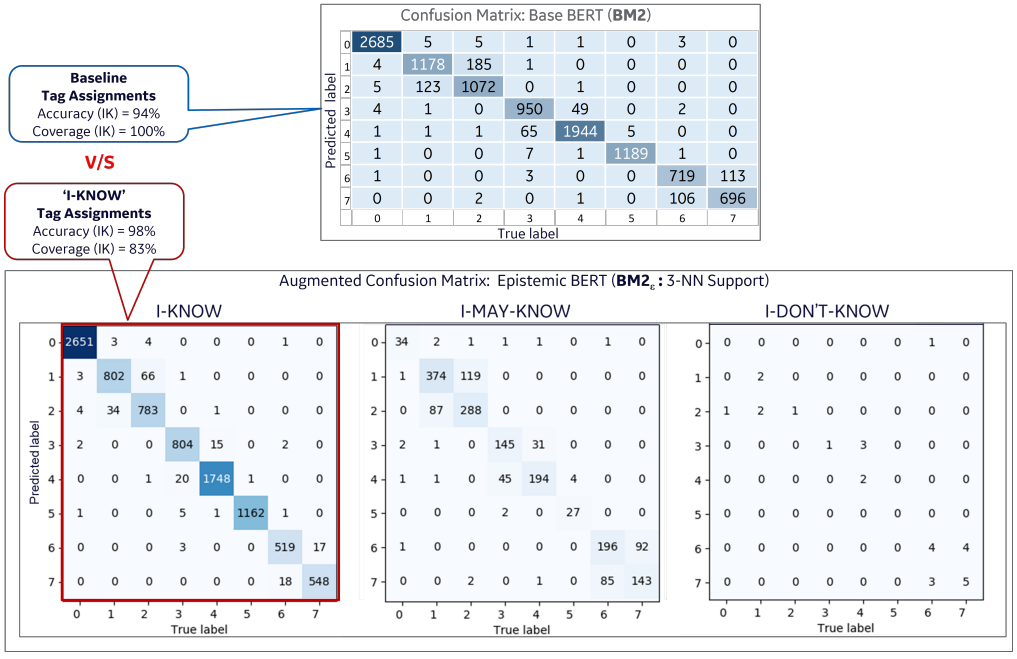

Figure 8. Comparing performance of the classifier for the case with fine tuning applied to the pre-trained BERT model, using the service records in the training data.

Base BERT		Epistemic BERT		
Model	Base Accuracy	Support	Coverage (F_{IK})	Accuracy (A_{IK})
Pre trained model	80%	3-NN	51%	97%
		10-NN	38%	99.2%
Fine-tuned model	94%	3-NN	83%	98%
		10-NN	73%	99%

Figure 9. Overall comparison of models under varying ablation conditions: base versus epistemic, pre-trained versus fine-tuned and 3NN support versus 10NN support.

recall and F1-scores. As before, the metrics for the Epistemic classifiers apply only to the *I-Know* matrix. As shown, the Epistemic BERT classifier with fine-tuning allows for inference with highest Precision, Recall and F1-scores for all 8 classes.

4.1. Impact

The high values of *coverage* (see Figure 9) for which the prediction reliability of the model is high indicates the amount of manual labor that can be avoided, in an otherwise manually tagged workflow. For one use-case considered within the Avation industry, it was estimated that for a fleet size of 300 aircraft, over 1,000 MX records are created daily with free text to describe issue and corrective action, making it a very labor-intensive process. An unfortunate by-product of the high volume of records to be manually tagged is the impact on the accuracy (seen to be about 90%) of the tagging from cognitive and information overload and the fatigue from it. Incorrect tags lead to additional labor-intensive effort to correct them, where in a single engineer can take upto 15 days to rectify 10k records. Similarly in nuclear domain, maintenance workorder planning is a time consuming process. An automated tagging system has potential to significantly reduce manual effort and partially automate planning based on existing procedures if predicted issue categories can be automatically identified (tagged) with high reliability. Our mixed initiative system shows that a fully automated tagging engine can replace the human engineer for more than 50% of the time, for which the prediction accuracy of the engine is expected to be greater than 97%. A critical by-product of reducing the manual effort is the consequent reduction in cognitive load for the human experts, thereby impacting their own tagging accuracy in a positive direction (this is a hypothesis based on how this has been seen to be true in domains like radiology using computer-aided diagnosis). As a result, we believe that the incorporation of mixed initiative tagging using the epistemic framework is expected to be highly relevant and valuable for regulated domains like nuclear, which are constrained by requirements of safety when it comes to decisions made without humans-in-the-loop, but also highly burdened by operational and maintenance (O&M) costs, thus driving up the costs of energy production. A mixed initiative system that logically balances model-risk from full automation with high O&M costs from heavy manual labor provides a pathway to the feasible introduction of high performance machine learning models, without incurring all of their risks. An additional burden that is brought upon machine learning model is the lack of transparency of their working, leading to a reluctance in their use in a fully automated mode. With the epistemic framework, the ability to present the alternatives in the training data that act a *support* for a highly reliable prediction act as an explainability characteristic for the model (i.e., *the model assigns a tag 'T' to service record 's' because here are 'k' other service records that share the same tag with the prediction*). This also permits the model to be audited on an individual prediction basis, allowing a visual comparison of the content in the service record 's' with the content in the 'k' service records that were identified as its neighbors in the model's latent spaces. From a decision support perspective, the tag predictions with epistemic status of *I-May-Know* can be used to further cue the human tagging expert with the candidate tags that might be most applicable to the service record at hand, using the subset of tags identified in the supporting training data records for the service record. As a result, the human expert does not have to consider the entire universe of tags to assign from, when brought in to intervene, and can get supplied with the most likely choice of tag assignments, which are presented to them for their consideration, which is further expected to improve their own performance. In nuclear domain context such a system can be utilized in a number of different ways, such as for classifying the issue based on described symptoms, identify relevant maintenance procedures/documents, workorder templates, and generating action recommendations for customers.

5. RELATED WORK

While BERT-based models are used extensively in general, their application to the analysis of Maintenance and Service records is fairly recent - (Usuga-Cadavid, Lamouri, Grabot, & Fortin, 2021) explore the use of two recent deep learning models (CamemBERT and FlauBERT) for natural language processing (NLP) to analyze unstructured data from maintenance logs. Primarily, they make use of LIME (Ribeiro, Singh, & Guestrin, 2016) on top of their models to address the issue of limited interpretability of ML models, that results in human reluctance when accepting model predictions. (Usuga-Cadavid, Grabot, Lamouri, & Fortin, 2021) use an alternate language model called GPT-2 to generate artificial maintenance reports, which are then employed to mitigate the class imbalance when training a Deep Learning (DL) model for analyzing unstructured data in maintenance logs. (Stenström, Al-Jumaili, & Parida, 2015) demonstrate the use of natural language processing (NLP), to make the process of manual analysis of maintenance records more efficient by

CLASS	BASELINE: No finetuning (for all samples)			BASELINE: finetuning (for all samples)			EPISTEMIC BERT: No finetuning (for 51% of the samples)			EPISTEMIC BERT: finetuning (for 83% of the samples)		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
a1	0.8147	0.8854	0.8486	0.9944	0.9941	0.9943	0.9811	0.9938	0.9874	**0.9970**	**0.9962**	**0.9966**
a2	0.6323	0.6618	0.6467	0.8611	0.9006	0.8804	0.9040	0.9150	0.9095	**0.9197**	**0.9559**	**0.9375**
a3	0.6806	0.5672	0.6187	0.8926	0.8474	0.8694	0.9429	0.8919	0.9167	**0.9526**	**0.9169**	**0.9344**
a4	0.7449	0.7057	0.7247	0.9443	0.9250	0.9346	0.9226	0.8746	0.8980	**0.9769**	**0.9652**	**0.9710**
a5	0.9262	0.9183	0.9223	0.9638	0.9735	0.9686	0.9777	0.9821	0.9799	**0.9876**	**0.9904**	**0.9890**
a6	0.9696	0.9578	0.9636	0.9917	0.9958	0.9937	0.9918	0.9927	0.9922	**0.9932**	**0.9991**	**0.9961**
a7	0.7054	0.6418	0.6721	0.8600	0.8652	0.8626	0.9547	0.9336	0.9440	**0.9629**	**0.9593**	**0.9611**
a8	0.6950	0.7491	0.7211	0.8646	0.8603	0.8625	0.9517	0.9654	0.9585	**0.9682**	**0.9699**	**0.9691**

Figure 10. Class-wise comparison of the 4 models along precision, recall and F1-scores (bold numbers indicate maximum values for the row).

relating text entry field data to other data fields. (Öztürk, Solak, Bäcker, Weiss, & Wegener, 2022) show use of maintenance reports as an augmented source of information related to a component's or a plant's operating status, indicating that information contained in maintenance reports can enhance the performance of data-driven models of prediction. They leverage state-of-the-art AI methods, including BERT based models, for the analysis of maintenance reports, to construct embeddings and clusters within them to identify groups of similar events that can help aid predictive maintenance tasks. (Saetia, Lukens, Pijcke, & Hu, 2019) describe a data-driven approach that employs machine learning-based and rule-based methods within a hybrid man-in-the-loop workflow for identifying equipment taxonomy from equipment records in maintenance management systems. (Lowenmark, Taal, Nivre, Liwicki, & Sandin, 2022) describe work that is also closely related to the work described in this paper, where the authors deal with the problem of how BERT-based models get impacted due to the occurrence of out-of-vocabulary (OOV) words, by learning to substitute technical terms with natural language descriptions. The work described in this paper can also contribute to the domains of maintenance records and technical language development. It also motivates exploration into the application of BERT representations, performance measurement, and frameworks that can act as suggestion engines for labeling data.

6. CONCLUSION

The outcomes from our experiments clearly indicate that justification-based (or epistemic) measures of reliability derived for individual predictions of machine learning models can greatly enhance the performance of an otherwise automated workflow that uses a machine learning classifier model for ATA tag assignment. Specifically, we showed that these epistemic measures provide a natural mechanism to design a mixed initiative system for ATA tagging of service records, such that predictions that signal high levels of epistemic certainty can be left unvetted, while the rest are channeled to human tagging experts. Our results demonstrate that this decomposition of labor within the mixed initiative system is effective because the high reliability tag assignments also show extremely high prediction accuracy. Conversely, tag assignments where the prediction reliability is found to be low (*I-May-Know* or *I-Dont-Know*) can either arise from class-confusion due to the specific content of the corresponding service records. These might be indicative of cases where there is genuine ambiguity with regard to the right ATA tag to assign, and where there is additional context that human experts would be more suited to handle, than a model. Low prediction reliability detected by the epistemic framework can also arise due to the content in the service record being novel or previously unseen, which is again better handled by a human expert than a model. For many AI systems being developed in the world today, while their statistical performance have been shown to superior across a large class of domains, their adoption has been curtailed due to considerations of transparency, along with domain-based requirements of assurance and safety. A mixed initiative system that assigns decision-making, on a sample-by-sample basis, between the AI model and a human expert, based directly on considerations of the model's prediction reliability for the sample, optimally addresses these barriers of adoption and ushers in a new era where computers and humans can seamlessly divide their decision-making as a direct function of their primary competences. This will enable the benefits of AI-based models to finally apply in larger scales across diverse industries.

Future Work

As continuation of this work several research questions are planned as part of future work. We are looking into scenarios where tagging categories may exhibit overlap, i.e. a maintenance record may belong to more than a single label. Furthermore, original data labeling is understood to contain inaccuracies or noisy labels. We intend to characterize the

effect of such inaccuracies and consequently robustness to noise using our prediction reliability approach. Additionally, these system of classifications can often be described through a system-subsystem-component hierarchy. Therefore, we intend to demonstrate use of humble AI-based prediction reliability in a hierarchical settings. For instance, it may be possible to make a prediction with high reliability only to a certain level in the hierarchy, and for a deeper level classification ambiguity may still exist. In such cases we expect high reliability predicted level to provide helpful context for further classification, thereby reducing operator burden and improving classification speed. As further extension to hierarchical classification we plan to include larger number of tag classes into our scheme (Seale et al., 2019), esp. recognizing the challenge from high imbalance in class distribution.

ACKNOWLEDGMENT

Research funding for this work was provided by Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0001290 towards advancing the use of AI for reducing O&M costs for nuclear power plants. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. The information and data presented herein were supported in part by GE Aviation in a prior collaboration. This work is a part of GE's Humble AI initiative.

REFERENCES

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, *29*(6), 141–142.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2019, minneapolis, mn, usa, june 2-7, 2019, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/n19-1423 doi: 10.18653/v1/n19-1423

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. Retrieved from https://github.com/google-research/bert

FAA. (2001). *Aviation maintenance alerts* (Vol. 43-16A; Tech. Rep. No. Advisory Circular). US Department of Transportation.

GEMINA. (2019). Generating electricity managed by intelligent nuclear assets. Retrieved from https://arpa-e.energy.gov/technologies/programs/gemina

GEMINA. (2020). AI enabled predictive maintenance digital twins for advanced nuclear reactors. Retrieved from https://arpa-e.energy.gov/technologies/projects/ai-enabled-predictive-maintenance-digital-twins-advanced-nuclear-reactors

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Acl*.

Lowenmark, K., Taal, C., Nivre, J., Liwicki, M., & Sandin, F. (2022). Processing of condition monitoring annotations with bert and technical language substitution: A case study. In *Phm society european conference* (Vol. 7, pp. 306–314).

Min, B., Ross, H. H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., ... Roth, D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey. *ArXiv*, *abs/2111.01243*.

Öztürk, E., Solak, A., Bäcker, D., Weiss, L., & Wegener, K. (2022). Analysis and relevance of service reports to extend predictive maintenance of large-scale plants. *Procedia CIRP*, *107*, 1551–1558.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Naacl*.

Radford, A., & Narasimhan, K. (2018). Improving language understanding by generative pre-training..

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners..

Ribeiro, M. 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* (pp. 1135–1144).

Saetia, K., Lukens, S., Pijcke, E., & Hu, X. (2019). Data-driven approach to equipment taxonomy classification. *Annual Conference of the PHM Society*.

Seale, M., Hines, A., Nabholz, G., Ruvinsky, A., Eslinger, O., Rigoni, N., & Vega-Maisonet, L. (2019). Approaches for using machine learning algorithms with large label sets for rotorcraft maintenance. In *2019 ieee aerospace conference* (p. 1-8). doi: 10.1109/AERO.2019.8742027

Stenström, C., Al-Jumaili, M., & Parida, A. (2015). Natural language processing of maintenance records data. *International Journal of COMADEM*, *18*(2), 33–37.

Steup, M. (2007). The analysis of knowledge. *Stanford encyclopedia of philosophy*.

Usuga-Cadavid, J. P., Grabot, B., Lamouri, S., & Fortin, A.

(2021). Artificial data generation with language models for imbalanced classification in maintenance. In *International workshop on service orientation in holonic and multi-agent manufacturing* (pp. 57–68).

Usuga-Cadavid, J. P., Lamouri, S., Grabot, B., & Fortin, A. (2021). Using deep learning to value free-form text data for predictive maintenance. *International Journal of Production Research*, *0*(0), 1-28. doi: 10.1080/00207543.2021.1951868

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Virani, N., Iyer, N., & Yang, Z. (2020). Justification-based reliability in machine learning. *ArXiv*, *abs/1911.07391*.

BIOGRAPHIES

Naresh S. Iyer Dr. Naresh Iyer is a Principal Scientist in the AI and Machine Learning (ML) group at GE Research. He has 22 years of experience in the research and application of machine learning to a variety of industry problems, including asset life prognostics, multi-objective optimization and decision making under uncertainty. He has developed solutions for a diverse range of industrial PHM applications using methods in supervised, unsupervised, semi-supervised learning and evolutionary soft computing. Recently, he was a lead contributor in an ARPA-E program targeting AI-based generative design to improve manufacturability of additively manufactured parts. He is currently executing on ARPA-E's GEMINA program dealing with the development of reliable machine learning models for predictive maintenance of nuclear reactors, and another program with DOE-AMO targeting the application machine learning for efficient inspection of large-scale additively manufactured aerostructure parts using DED. His research interests include reliable machine learning, sequential decision making, and adversarial machine learning. He has over 30 publications and 45 patent filings. He has served in multiple program committees for international conferences and journals.

Nurali Virani Dr. Nurali Virani is a Senior Scientist in the Machine Learning team at GE Research. He is a multidisciplinary researcher with a strong academic and research background in machine learning, surrogate modeling, control theory, sensor fusion, signal processing, and motion planning. He has worked on several projects including AI-driven control of wind farms, AI-driven safe control of power generation gas turbine units, predictive maintenance digital twins for industrial assets, and characterizing robustness and trust in ML models. His current research interest is in making AI aware of its competence and to improve its competence and robustness via continuous learning (Humble AI). He was awarded GE Global Research CTO Technology Award (5 Under 5) for Outstanding Research in 2018 as well as 2019 Rudolph Kalman Best Paper Award by ASME. Prior to joining GE, he was a research assistant at Penn State. He was awarded a silver medal for academic excellence, when he graduated from Indian Institute of Technology Kharagpur in 2011. Dr. Virani has 30+ peer-reviewed publications and 6 patents.

Zhaoyuan Yang Mr. Zhaoyuan Yang is a research engineer in the Computer Vision group at GE Research. Zhaoyuan has developed multiple ML/CV-based solutions for industrial inspections as well as security of cyber-physical systems. His research interests include robustness, trustworthiness as well as explainability in machine learning models.

Abhinav Saxena Dr. Abhinav Saxena is a Principal Scientist in AI & Learning Systems at GE Research. Abhinav has been developing ML/AI-based PHM solutions for various industrial systems (aviation, nuclear, power, and healthcare) at GE and has been driving integration of AI-based PHM analytics in GE's industrial systems. He is the PI for ARPA-E GEMINA program developing AI-Enabled Predictive Maintenance Digital twins for Advanced Nuclear Reactors. Abhinav is also an adjunct professor in the Division of Operation and Maintenance Engineering at Luleå University of Technology, Sweden. Prior to GE, he was a Research Scientist with SGT Inc. at NASA Ames Research Center for over seven years. Abhinav's interests lie in developing PHM methods and algorithms with special emphasis on deep learning and data-driven methods in general for practical prognostics. He has published over 100 peer reviewed technical papers and has co-authored a seminal book on prognostics. He actively participates in several SAE standards committees, IEEE prognostics standards committee, and various PHM Society educational activities, and is a Fellow of the PHM Society. He also served as chief editor of International Journal of Prognostics and Health Management between 2011-2020.