Evaluating Vector Representations of Short Text Data for Automating Recommendations of Maintenance Cases

Akshay Peshave[1], Kareem S. Aggour[1], Asma Ali[2], Varish Mulwad[1], Sharad Dixit[1], and Abhinav Saxena[1]

[1] *GE Research, Niskayuna, New York, 12309, USA*
akshay.peshave@ge.com, aggour@ge.com, varish.mulwad@ge.com, sharad.dixit@ge.com, asaxena@ge.com

[2] *GE Digital, Chicago, Illinois, 60661, USA*
asma.ali1@ge.com

ABSTRACT

Nuclear power is a carbon-free source of energy, and features as a key component in the mix of energy towards meeting ambitious decarbonization goals. However, as it currently stands, nuclear power generation is orders of magnitude more expensive when compared to fossil energy sources. Recently, there has been a significant push, by both the US government and the power industry, to identify and address opportunities for cost reductions in nuclear power generation. While capital costs are being addressed through innovative Smart Modular Reactor (SMR) designs, reductions in Operations and Maintenance (O&M) costs is an equally important opportunity.

Enhanced methods of remote monitoring and health management-based maintenance optimization have been recognized as key elements to reduce preventive and corrective maintenance costs. Beyond sensor data and signal processing, there is also significant potential for reducing manual efforts and improving productivity in maintenance scheduling and planning activities. We can accelerate maintenance processes and reduce human effort by automating maintenance recommendation generation using state-of-the-art Technical Language Processing (TLP) methods that analyze large volumes of historical maintenance case data.

This paper presents efforts towards developing a prescriptive maintenance system that integrates with and enhances commercially available state-of-the-art asset performance management software. The goal of prescriptive maintenance is to analyze the behavior of an asset, assess its condition, and recommend specific actions to maximize the utility of that asset. Specifically, this work evaluates three approaches of different complexities for vectorization of short-text maintenance case titles for kNN-based recommendation of cases relevant to a new input case title. Industrial text must first be vectorized to build automated and/or machine learning-based prediction and recommendation models. The choice of vectorization methods heavily dictates how the language gets modeled and consequently impacts the performance of downstream prediction and recommendation models.

The objective of the nearest neighbor case recommendations is to reduce manual Subject Matter Expert (SME) effort and increase consistency of recommended maintenance actions on industrial assets by reusing actions performed on the identified nearest neighbor cases from past maintenance work. Four models based on three text vectorization approaches are evaluated, quantitatively as well as qualitatively, using real world data from a large variety of utility customers in the energy domain. A single tier (WVEC-1tier) and a three-tier (WVEC-3tier) approach that represent case titles in word-based vector spaces each significantly outperform a more complex bag-of-phrases topic vector space-based approach (TVEC-$\mathcal{K}$topics). We present our findings and challenges identified so far in building such a recommendation system.

1. INTRODUCTION

Analysis shows that in the near future, and likely even in the longer term, a sustainable decarbonization strategy will include a significant mix of renewable and nuclear sources of power. Nuclear energy in particular will be key to maintaining a constant supply of power during weather conditions that are adverse to solar and other renewable sources, especially in regions such as northern Europe and Canada.

The current cost of nuclear energy is substantially higher than fossil and most other renewable sources of energy, even after discounting for carbon credits. Capital and operational expenditures account for a large portion of that cost. While capital costs can be reduced by Advanced Reactor (AR) designs, such as the BWRX-300 currently under development by GE Hitachi, operational costs must also be reduced by

keeping these plants well-maintained and efficient throughout their operational life. Several efforts are underway to develop robust methods for generating large volumes of sensor-based measurement data for remote monitoring and semi-automated health management of nuclear assets. However, significant benefits can also be realized by improving the troubleshooting and maintenance planning processes of these assets as well.

1.1. Prescriptive Maintenance

The goal of prescriptive maintenance within Prognostics and Health Management (PHM) is to analyze the behavior of an asset, assess its condition, and recommend specific actions to maximize the utility of that asset. The recommendations must balance the benefits and implications of keeping the asset running, even in a partially degraded state, with those of an unplanned shutdown for repair. This is a goal across industries including nuclear and non-nuclear power generation, oil & gas, and mining.

Learning from past resolutions and automating analysis of incoming cases to generate recommendations is an ongoing effort in PHM. While a large number of analytical models are employed for heath assessment and alert generation, the disposition of alerts and subsequent maintenance decision-making is performed largely manually at Remote Monitoring Centers (RMCs). Given the manual nature of the maintenance recommendation process, the lead time, quality and consistency of maintenance recommendations depends on the effectiveness of the analytical models, volume of alerts, and field engineers' experience.

1.2. Perceived Business Value

Depending on the nature of an alert, its sensitivity, and how early it occurs in the P-F (potential failure) curve, PHM decisions dictate the business value. Business value is driven by a combination of early detection and speedy decision-making to choose one or more requisite preventive and corrective actions that can be completed within the available time.

Ideally, PHM processes must allow RMCs to consistently and efficiently generate effective recommendations that enable the plant reliability teams to review, diagnose, and plan mitigation & corrective actions (e.g. labor, material, costs, scheduling) accordingly. These assessments are often performed manually by SMEs from various domains. The SMEs analyze time series and alert data from sensors, document their findings, and generate recommended actions that feed into work orders. This largely manual and effort-intensive process results in semi-structured and unstructured text data being generated for every maintenance 'case' describing the observations, suspected conditions, and recommendations.

Automating the generation of the recommended actions based on unstructured text that summarizes the sensor data observations would reduce SME effort in the PHM process and accelerate the time to maintenance decisions. The value proposition becomes even more attractive for fleet RMCs with thousands of assets being monitored in real time. Such automation of maintenance action recommendations will allow RMCs to produce quality and consistent maintenance recommendations for plants regardless of the analysts' level of experience.

1.3. Approach

This paper addresses the task of extracting a list of relevant past cases ranked by their similarity to the input case. Further, the complexity of the task is increased by attempting to train the model using only case titles, which are short text data that are sparse, coarse-grained summaries of the case details. In this work, we evaluate three approaches to vectorize case titles and automatically cluster them for modeling k nearest neighbor classifiers to identify relevant past cases for a new case title. The intent is to integrate the model into an online prescriptive recommendation process to rapidly generate a ranked list of nearest neighbors for each new case in order to enable SMEs to more efficiently recommend actions to address the case.

Further, we seek to identify future directions of research towards our long-term objective of using refined Technical Language Processing (TLP) on large volumes of industrial data to automatically classify new cases and provide consistent, quality recommended actions in order to gain efficiencies by accelerating the time to maintenance decisions.

This work is intended for PHM of nuclear power-related assets. Due to the sensitivity of data in that domain and consequent data acquisition challenges, we use similar data from non-nuclear power generation equipment provided by GE Digital. Since the "balance of plant (BOP)" equipment is similar in both domains and our approach is generalized, non-nuclear power domain data has been used to evaluate our approach in this paper. With appropriate domain dictionaries and taxonomy, findings presented in this work apply to data in the nuclear power domain without loss of generality.

The remainder of this paper is organized as follows. Section 2 summarizes related prior art and outlines our core contributions. Section 3 briefly describes GE Digital's Asset Performance Management (APM) solution as the source of data for this research, followed by a description of the evaluation data in Section 4. Section 5 outlines our modeling architecture and the different modeling approaches followed by an evaluation and discussion of results in Section 6. Finally, Section 7 summarizes our findings and outlines future research directions.

2. PRIOR ART

Text pre-processing, in the form of stemming, stop word removal, and vocabulary-based normalization has been previ-

ously explored for PHM-related use cases. Many approaches have utilized these in conjunction with word representation models for TLP (Nandyala, Lukens, Rathod, & Agarwal, 2021). A few examples in the maintenance and support domains include (Edwards, Zatorsky, & Nayak, 2008; Salo, McMillan, & Connor, 2019; Hansen, Coleman, Zhang, & Seale, 2020; Sexton, Brundage, Hodkiewicz, & Smoker, 2018). Approaches that utilize phrase-based TLP for various use cases have typically relied on n-gramming of text or pre-defined vocabularies of concept tags (Navinchandran, Sharp, Brundage, & Sexton, 2019; Pau, Tarquini, Iannitelli, & Allegorico, 2021; Ottermo, Håbrekke, Hauge, & Bodsberg, 2021).

Prior work that explored the clustering of maintenance cases include: (Pau et al., 2021), which generated word embedding models for clustering cases for recommendations; (Salo et al., 2019), which evaluated an SME-in-the-loop, iterative clustering approach for work orders to identify high frequency events; and (Edwards et al., 2008), which used singular value decomposition on TF-IDF vectors to reduce the dimensionality of the vector space to cluster cases for classification.

(Salo et al., 2019) clustered historical corrective maintenance work orders to identify high frequency events. Theirs is an active learning scheme that involved a two-stage clustering approach with a human in the loop to review the clusters and provide feedback to help guide and retrain the algorithm, reducing the original clustering time from 10 days to 1 day, 6 hours of which was the manual annotation time, thus requiring manual intervention to achieve a significant performance improvement.

(Pau et al., 2021) tackled a very similar problem to our own and provided some insightful observations about the different approaches they explored, but they performed TLP on the full case text including event descriptions, technical assessments, and recommended actions, whereas we are only clustering on the case titles, making our volume of text significantly leaner.

In the prior art we find one of two scenarios (and in many instances, both scenarios). We find the related work is either only partially automated and thus requires manual intervention to facilitate the case clustering, such as in (Salo et al., 2019; Ottermo et al., 2021), and/or the related work is processing a significant volume of text from each case to perform the clustering, such as in (Hansen et al., 2020; Pau et al., 2021).

Our primary contribution is we demonstrate that maintenance cases can be clustered with reasonably high quality (i) using an entirely automated pipeline, and (ii) with only a small fraction of unstructured text by relying solely upon the short case titles. These contributions are important because (i) we want to completely automate the case clustering towards our long-term goal of producing a completely automated maintenance recommendation generation pipeline, and (ii) relying upon the case titles alone is desirable because it allows the performance to be independent of the quality of the case details written by the SMEs. We see significant variability in the breadth and depth of detail written in the case text across different engineers and especially across different sites, and by focusing on the case titles alone we avoid challenges associated with this significant variability in text quality and quantity.

3. ASSET PERFORMANCE MANAGEMENT SYSTEM

GE Digital's Industrial Managed Services (IMS) team's Asset Performance Management (APM) software is a predictive and preventative maintenance software system that employs AI/ML models, analytics, and digital twins for a wide range of physical assets. It enables reliability and maintenance engineers to detect early warning indicators and preempt equipment failures by providing features such as monitoring, issue triage, customer notification, remediation guidance, analytics management, and strategy management for assets.

Over the years, IMS has used APM to monitor thousands of assets around the globe for many dozens of customers serving verticals such as power, nuclear, oil & gas, mining, aviation, chemicals, and more.

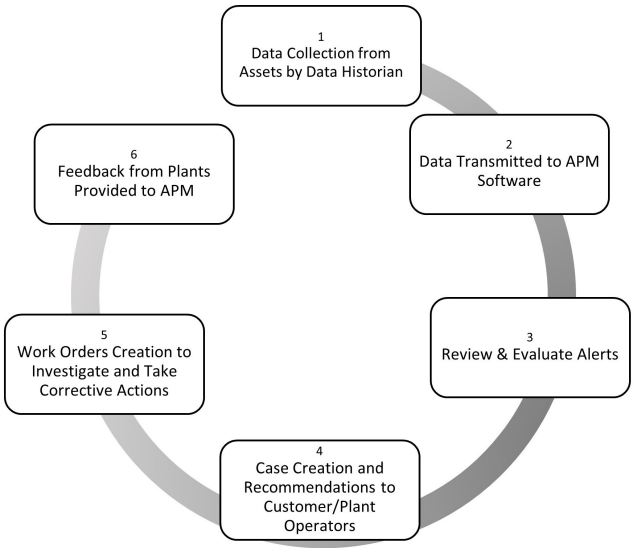


Figure 1. Asset Performance Management (APM) workflow

The APM workflow shown in fig. 1, results in the creation and analysis of a few different types of data. In Step 1, data is collected online and offline from assets directly into the onsite "historian", the time series data store that collects both times series sensor data and alerts. In Step 2, this data is transmitted to GE Digital's APM software in real time.

In Step 3, the transmitted data is processed via analytical models that are based on each individual asset's historical

data and compared in real time to actual data against normal and predicted behavior for all sensors from that asset. Any anomalies in the data are then compared to precursor signatures based upon known fault patterns and persisted. APM detects and identifies unexpected events and abnormal behaviors and formalizes those into one or more alerts and notifications. These alerts and associated data patterns are then reviewed and analyzed by the SMEs/analysts in the APM monitoring center in Step 4, who then make the determination to generate a new case and make a recommendation to the customer. These cases include a title, textual description of the observations, possible reasons for the change in asset performance, and potential recommended actions for the customer. This step requires significant time investment and expertise to effectively diagnose potential issues indicated by the alerts and to generate corresponding recommendations for the customers. The key hypothesis of this research revolves around being able to learn these recommendations from past cases based on short case titles and then provide the SMEs with a subset of plausible recommended actions in near real time. The case titles generated at this step are used as input to the modeling pipeline described in this work.

In Step 5 of the workflow, customers at the plant would then generate a work order to review, investigate, and perform any corrective actions. Finally, in Step 6, feedback is sent back to APM to inform if the recommended action helped the customer get to the root cause of the problem, mitigate, and disposition the issue correctly.

A case is generated when alerts are deemed actionable. The case details are reviewed and then work orders are generated for onsite teams to diagnose and perform corrective actions on the impeding asset and/or process. This work utilizes a large set of historical case data from APM related to the power domain.

4. DATASET

Case titles from a sample of data collected over 5+ years by GE Digital in their Asset Performance Management (APM) software are utilized for evaluation of the models. The data sample contains 20,257 maintenance cases of which 19,570 cases titles are in English and are not marked as duplicate or deleted in the APM system. This set of 19,570 cases are split into a training set containing 15,657 (80%) cases and a held-out test set of 3,913 (20%) cases. Basic data statistics about the number of unique words and noun-phrases for the entire dataset as well as the training and test sets are provided in table 1.

The test set case titles contain terms (words or noun-phrases) that are not observed in the training set case titles. This emulates real-world PHM workflows where new textual terms are observed as new cases are input to the systems when they are operational. The test set case titles consist of 1,590 unique words and 1,867 unique noun-phrases of which ≈75% and ≈67%, respectively, are observed in the training set. The terms that are unobserved in the training set are not considered when characterizing the case titles for recommending related cases as is explained in section 5.

Dataset Split	# Cases	# Words	# Noun-phrases
All	19,570	3,662	4,852
Train	15,657	3,270	4,233
Test	3,913	1,590	1,867

Table 1. APM Dataset Basic Statistics

Additionally, a domain-specific dictionary is utilized that maps 1,945 abbreviations and shorthand terms to 863 normalized terms. This dictionary contains abbreviation mappings such as "HRSG"→"heat recovery steam generator" and shorthand mappings such as "aux xfrmr"→"auxiliary transformer".

5. METHOD

The task addressed in this work is that of identifying cases from the training set that are related to a new case in order to utilize the knowledge of previously performed actions from the related cases as candidates for recommended actions or next steps for the new case. Further, we operate under the constraint of relying solely on the case titles for the purpose of identifying similar cases. Our modeling pipeline involves case title text pre-processing, title vectorization, title clustering, and k-nearest neighbor classification as the core stages. This pipeline is shown in fig. 2.

We further compare the effectiveness of projecting case titles in word-based and topic-based vector spaces for vectorization, with the word-based vector space being employed in a single tier and three tier fashion. The internal workflows of these vectorization approaches is shown in fig. 4. The word-based, tiered vector space is relatively more nuanced and this is reflected in the clustering and kNN model stages of the pipeline in fig. 2. The subsections that follow describe each stage of the modeling pipeline in detail.

5.1. Text Pre-Processing

Our text pre-processing stage involves treatment of the case titles to prepare them for the downstream stages. The pre-processing steps are as follows:

1. **Noun-phrase Extraction**: This step performs noun-phrase extraction from case titles. Noun-phrases are nouns (i.e. subjects and objects) in a natural language sentence along with their corresponding descriptor words (i.e. articles and adjectives). They reasonably represent domain-specific context in the form of components (e.g.,

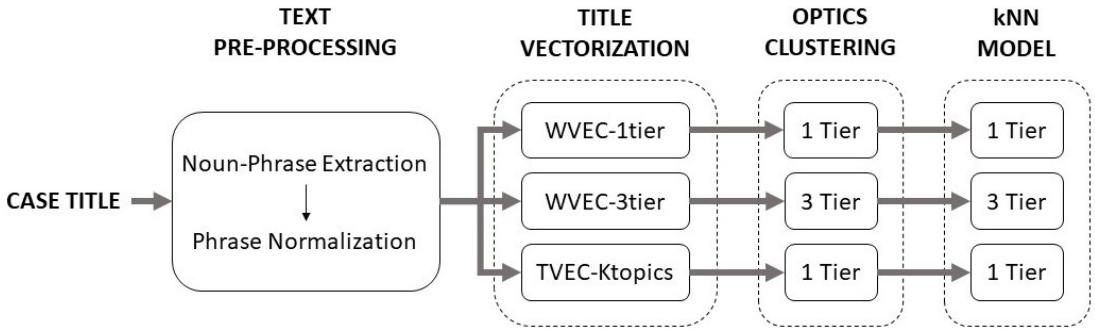


Figure 2. Modeling Pipeline

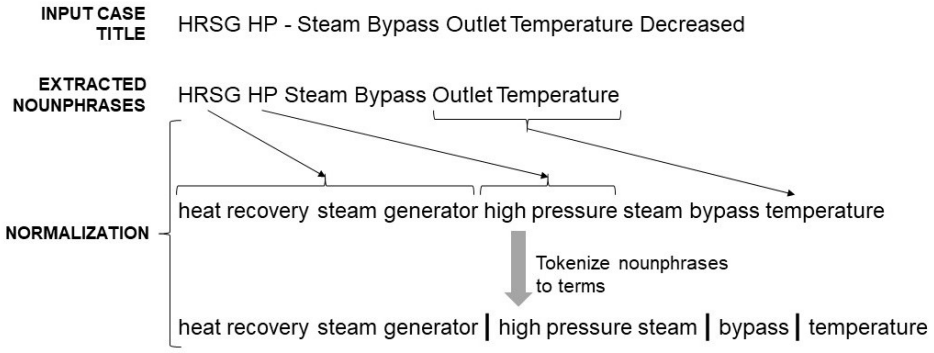


Figure 3. Text Pre-processing Example

"gas turbine", "fan blade", "compressor") and conditions (e.g., "increased vibrations", "sensor issue", "pressure swings") pertaining to the case. spaCy[1], a natural language processing toolkit, is utilized for this step. spaCy extracts noun-phrases without allowing for nesting of noun-phrases, which is important to our approach.
Verbs, and associated adverbs, if any, may also represent conditions, and possibly their directionality, in addition to establishing relationships between the aforementioned entities. The scope of this work is restricted to the utility of noun-phrases. Verbs are out-of-scope but addressed in the evaluation discussion to motivate future refinements.

2. **Phrase Normalization**: A domain-specific, curated dictionary (refer section 4) is used to normalize abbreviations and shorthand terms that are encountered in the extracted noun-phrases. Further, we utilize language-specific lemmatized forms of noun-phrases and disregard articles, if any, contained in the noun-phrases to ensure consistency of phrases across the textual data. Lastly, the noun-phrases are split when parts of them appear as terms in the domain dictionary.

Fig. 3 shows an example of the text pre-processing for a case title. The case title contains two abbreviations, "HRSG" and "HP". Their expanded forms in the dictionary are "Heat Recovery Steam Generator" and "high pressure", respectively. Further, "outlet temperature" is mapped to the term "temperature" in the dictionary. The title noun-phrases after normalization are split when dictionary terms appear as sub-phrases. You can see the split terms as the final output of text pre-processing in the figure.

[1]spaCy v3.2: https://spacy.io/

5.2. Inverse Document Frequency (IDF) Term Filtering

IDF is a measure of how frequent or infrequent a term appears across a document corpus. We use min-max scaled, log-IDF values in this work as defined below:

$$\widetilde{IDF}(v \in V) = \mathcal{S}\left(log\left(\frac{|D|}{|\{d : d \in D, v \in d\}|}\right)\right)$$

where, D is a training set of case titles, V is the vocabulary of terms in D and $\mathcal{S}(idf) : \mathbb{R} \rightarrow [0.01, 1]$ is the min-max scaling transform that scales log-IDF values for the vocabulary terms to the range [0.01,1].

This stage is used by each vectorization approach described in section 5.3 to filter the vocabulary of terms that are used to process case titles downstream. It filters out terms whose $\widetilde{IDF}$ values lie outside a pre-configured range. The intent of this step is to filter out terms in the vocabulary that:

1. Occur so infrequently as to be considered too fine-grained to characterize domain context represented by

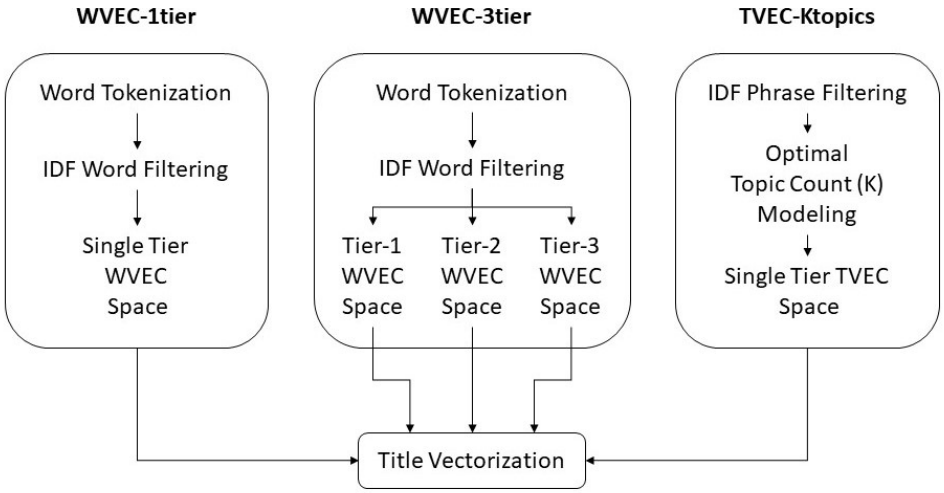


Figure 4. Title Vectorization Approaches

case titles. These are terms whose $\widetilde{IDF}$ is above the pre-configured $\widetilde{IDF}$ threshold.

2. Occur so frequently as to be considered too coarse-grained or broad to be discriminative of domain context represented by case titles. These are terms whose $\widetilde{IDF}$ is below the pre-configured $\widetilde{IDF}$ threshold.

5.3. Text Vectorization

In this work we evaluate three methods for vectorizing case titles. The methods differ in their complexity and the number of parameters they require for representing case titles in a vector space that feeds into the case clustering stage. Fig. 4 shows a visual description of the different vectorization approaches.

5.3.1. Single Tier Word Vector Space (WVEC-1tier)

This approach tokenizes normalized noun-phrases into word tokens followed by $\widetilde{IDF}$ filtration of the word tokens to create the WVEC-1tier vector space. The $\widetilde{IDF}$ filter retains words whose $\widetilde{IDF}$ values lie within a pre-defined range. This approach represents each case title (*ct*) as a vector $\vec{ct}$ of size $1 \times |\mathcal{V}'|$ where $\mathcal{V}$ is the vocabulary of word tokens, and $\mathcal{V}' \subset \mathcal{V}$ is the subset of word tokens in $\mathcal{V}$ that are retained by the filter. Each component of the vector $\vec{ct}$ is $\vec{ct}_{v \in \mathcal{V}'} = \mathit{freq}(v|ct)$, i.e. the frequency of occurrence of the word v in the title.

5.3.2. Three Tier Word Vector Space (WVEC-3tier)

This approach tokenizes normalized noun-phrases into word tokens followed by $\widetilde{IDF}$ filtration of word tokens that is the same as in section 5.3.1. Further, the filtered vocabulary $\mathcal{V}'$ is split into three non-intersecting sets of word tokens based on three pre-defined $\widetilde{IDF}$ ranges. This approach represents each case title as a vector $\vec{ct}^{\,i}$ of size $1 \times \sum_{a=1}^{i} |\mathcal{V}'_a|$ in each of the $1 \leq i \leq 3$ vector spaces where $\mathcal{V}'_i \subset \mathcal{V}'$ and $\mathcal{V}'_i \bigcap \mathcal{V}'_j = \varnothing$ for all $i, j \in \{1, 2, 3\}, i \neq j$. Each element of these vectors represents the frequency of occurrence of the corresponding word from the tier vocabulary in the title. Note that the $\vec{ct}^{\,3}$ in WVEC-3tier is equal to the vector $\vec{ct}$ in WVEC-1tier.

5.3.3. Single Tier Topic Vector Space (TVEC-$\mathcal{K}$topics)

This approach uses the normalized noun-phrases directly as tokens and performs $\widetilde{IDF}$ filtration on these tokens. A Latent Dirichlet Allocation (LDA) (Blei, Ng, Jordan, & Lafferty, 2003) topic model is trained to learn a topic mixture model for the corpus of case titles using the filtered tokens as the vocabulary. This work utilizes online LDA (Hoffman, Blei, & Bach, 2010) as implemented in *Gensim*[2].

An LDA topic model requires the topic count ($\mathcal{K}$) as a parameter. A topic count may be provided based on a priori knowledge of the dataset and/or the domain. In the absence of a priori knowledge, an optimal range of topic counts can be estimated by optimizing topic coherence (Röder, Both, & Hinneburg, 2015) over topics learned by the topic model for different topic counts.

The intuition for the TVEC method is to use the highest topic count that achieves near-maximum topic coherence among the candidate topic counts. This ensures near-optimal topic coherence with the highest possible resolution of the topic mixture model inferred for the case title corpus. This involves a manual process of inspecting topic coherence values. This approach of vectorization, and consequently this model, has more complexity and effort requirement than the WVEC models.

The TVEC approach represents each case title as a vector of size $1 \times \mathcal{K}$ in the topic vector space where $\mathcal{K}$, the topic count,

[2]Gensim v4.0 Mutlicore LDA: `https://radimrehurek.com/gensim/models/ldamulticore.html`

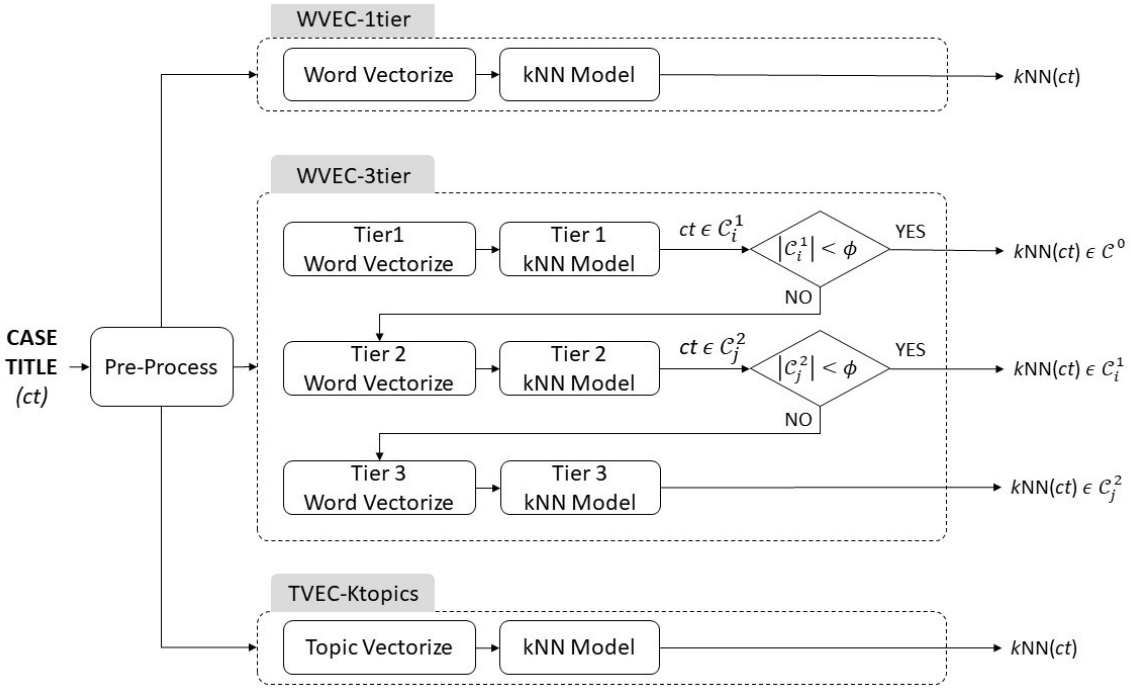


Figure 5. Recommendation Generation Workflow

is the dimensionality of the topic vector space. Each element of the vector is $\vec{ct}_k = Pr(topic_k|ct)$, the probability of *ct* belonging to the k^{th} topic.

5.4. Case Clustering

The case clustering stage of the pipeline uses the vectorized case titles to extract case clusters in an unsupervised manner. Extracting case title clusters enables our methods to utilize a k-nearest neighbor approach to recommend cases from the training set that are nearest neighbors of the input case title in the corresponding vector spaces.

Density-based clustering methods, such as DB-Scan (Ester, Kriegel, Sander, & Xu, 1996) and OPTICS (Ankerst, Breunig, Kriegel, & Sander, 1999), extract clusters of arbitrary shapes from the data based on data point density computation. DB-Scan requires a density parameter to be specified for clustering. OPTICS is a non-parametric density-based clustering approach that infers the density parameter and extracts clusters from the data by either plugging in the inferred parameter into DB-Scan or by using the 'Xi' parameter. This work uses the OPTICS implementation in scikit-learn[3] with cosine similarity as the distance metric and Xi-parameter clustering to extract case clusters.

OPTICS clustering of cases vectorized using WVEC-1tier and TVEC-$\mathcal{K}$topics is single tiered and, hence, straightforward. Clustering of cases using WVEC-3tier vectorization is done iteratively in three tiers. At each tier (t), $1 \leq t \leq 3$, sub-clusters are extracted for each cluster (C_i^{t-1}) from the previous tier if and only if $|C_i^{t-1}| \geq \phi$, where ϕ is a predefined cluster cardinality threshold and C^0, the initial cluster, is the complete set of case titles.

5.5. k-Nearest Neighbor (kNN) Classifiers

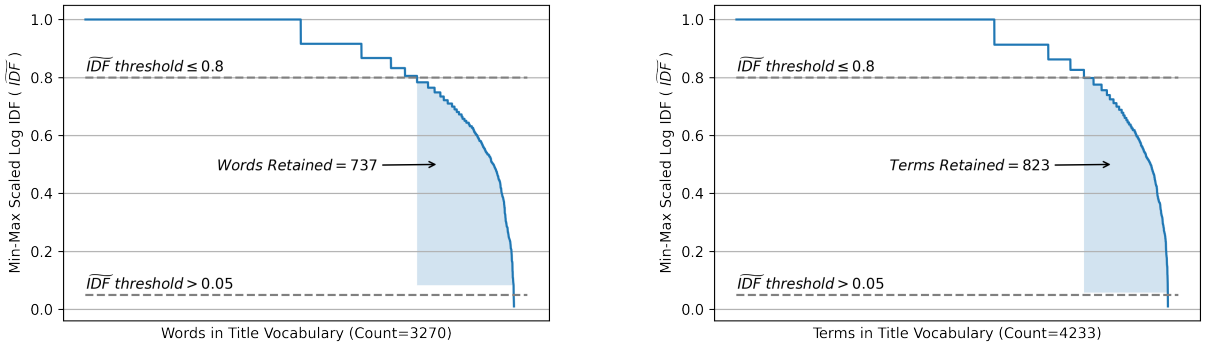
A k-nearest neighbor classifier model classifies an input data point to the majority class that its k-nearest neighbors from the training set belong to. This work trains kNN classifiers on clusters of case titles extracted using the different vectorization approaches described in section 5.3 using cosine similarity as the distance metric. The trained classifiers are then used to identify k-nearest neighbors for an input case as recommendations.

The WVEC-1tier and TVEC-$\mathcal{K}$topics approaches are single tiered and, hence, involve only one kNN classifier. The WVEC-3tier approach involves multiple kNN classifiers, $k\text{NN}_i^t$, corresponding to each tier-wise cluster, C_i^t, extracted during clustering.

5.6. Cases Recommendation Generation

The trained kNN classifier models are used to generate case recommendations from the training set for a new case title after it is pre-processed and vectorized. The recommendation generation workflows, with added nuance for the WVEC-3tier vectorization approach, are shown in fig. 5.

[3]scikit-learn v1.0 OPTICS: `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS.html`



(a) Distribution for Case Title Words used in WVEC

(b) Distribution for Case Title Terms/Phrases used in TVEC

Figure 6. $\widetilde{IDF}$ Distributions with Filtering Thresholds for Tokens used in WVEC and TVEC Vectorization of Case Titles

Case recommendation generation using the WVEC-1tier and TVEC-$\mathcal{K}$topics approaches involve word and topic vectorization, respectively, of the pre-processed case titles followed by nearest neighbor case identification by their corresponding kNN models. Recommendation generation for a case title ct using the WVEC-3tier approach involves identifying the finest-grained cluster $C_i^{\hat{t}}$ that the case title belongs to, where $\hat{t} = \arg\max_t \exists C_i^t : ct \in C_i^t$. The case title's nearest neighbor cases are then identified using the corresponding $k\text{NN}_i^{\hat{t}}$ classifier.

6. EVALUATION & DISCUSSION

The training set is used to train the kNN classifiers for the various vectorization approaches. Tokens in the test set case titles that are unobserved in the training set are disregarded when vectorizing test set case titles. All the vectorization approaches filter their respective token vocabulary to utilize only those tokens whose $\widetilde{IDF}$ lies in the (0.05, 0.80] range. The WVEC-3tier vectorization additionally segments its word vocabulary into three segments (or tiers) that include words whose $\widetilde{IDF}$ lies in the ranges (0.05, 0.25], (0.25, 0.50] and (0.50, 0.80], respectively.

Fig 6 shows the $\widetilde{IDF}$ distributions for the word vocabulary (fig. 6a) and term vocabulary (fig. 6b) used by the WVEC and TVEC vectorizations, respectively. The aforementioned $\widetilde{IDF}$ thresholds effectively filter out vocabulary tokens that are either too unique or too common amongst cases in the dataset. Further, the thresholds help reduce the dimensionality of the vector spaces to tokens that lie in the shaded regions in figs. 6a and 6b. The above predefined $\widetilde{IDF}$ ranges are effective for assessing the effectiveness of our models using the dataset at hand without burdening this paper with extensive optimization of the ranges, which is left for future work.

The minimum cluster size and minimum samples in the neighborhood of a core point for OPTICS clustering is set to 5 cases for all models. Cosine similarity is used as the distance metric for the WVEC and TVEC models.

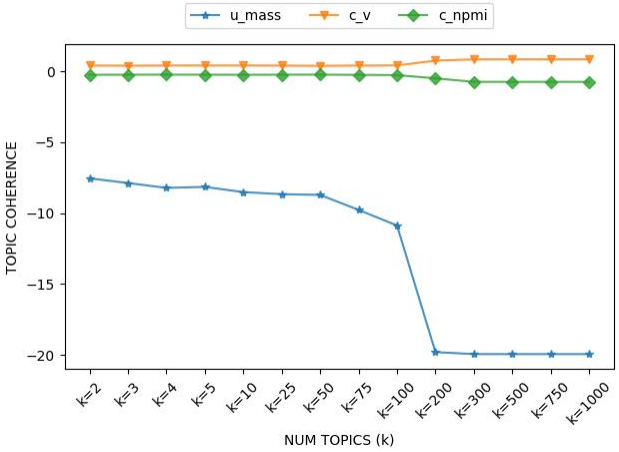


Figure 7. Topic Coherence of LDA Topic Models for Varying Numbers of Topics ($\mathcal{K}$)

The TVEC-$\mathcal{K}$topics approach requires manual inspection of topic coherence for different $\mathcal{K}$ values to choose the optimal topic count for training the LDA topic model. We evaluate topic coherence for LDA topic models on the training data with 14 different topic counts in the [2,1000] range as shown in fig. 7. We utilize three topic coherence metrics described in (Röder et al., 2015): (a) C_{UMass}, (b) C_v, and (c) C_{NPMI}. Fig 7 shows that the maximum $\mathcal{K}$ value that achieves topic coherence in close vicinity of the maximum observed value on all three topic coherence metrics is 50. For the purpose of this evaluation we utilize TVEC-25topics and TVEC-50topics models, with topic counts of 25 and 50 respectively, that have identical topic coherence. This is done to assess any difference in performance of the TVEC approach with different topic counts at the same topic coherence.

6.1. Quantitative Evaluation

The quantitative evaluation of the models is conducted using mean Jaccard similarity of test case titles with their corresponding neighbors at different ranks as an intrinsic quanti-

tative measure of performance. The Jaccard index is defined as the intersection over union of a set of items that represent a pair of points. In our case, the two data points being compared are the test case title (*ct*) and each individual neighboring case title (*nt*) identified for the test case title. The Jaccard index for the title pairs is measured by the intersection of word tokens over the union of word tokens present in the two titles after normalization:

$$Jaccard(ct, nt) = \frac{|\{w_i \in ct\} \bigcap \{w_j \in nt\}|}{|\{w_i \in ct\} \bigcup \{w_j \in nt\}|}$$

The Jaccard index value lies in the range [0,1]. A Jaccard index of 0 indicates there is no intersection of words between the two case titles being compared. A Jaccard index of 1 indicates the case titles being compared are comprised of identical sets of words.

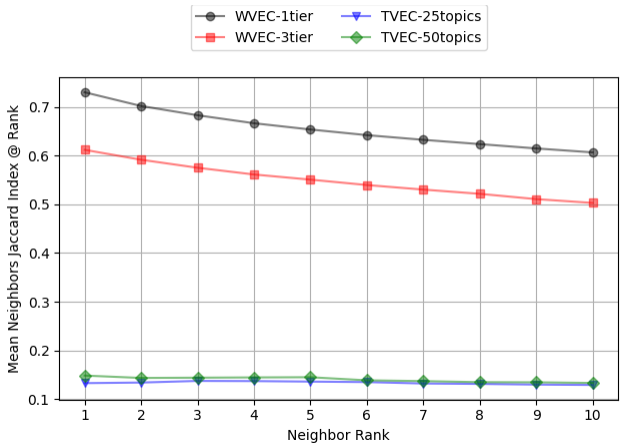


Figure 8. Mean Neighbor Jaccard Similarity for WVEC and TVEC Models at Increasing Neighbor Ranks

The mean neighbor Jaccard index for all models over the entire test set is plotted in fig. 8. Each point plotted at rank i in fig. 8 is the mean Jaccard similarity between all case titles in the test set and their corresponding neighbor at rank i identified by the respective model. WVEC-1tier consistently outperforms the other methods at all ranks. The TVEC models have comparable mean neighbor Jaccard similarity and identify neighbor case titles that are largely dissimilar to the input case title. This shows that the latent topic vector space based on noun-phrases is unable to represent case titles such that similar cases lie in close vicinity to each other and are separated from dissimilar case titles.

While WVEC-3tier underperforms WVEC-1tier, the parameters for creating the three-tiered word vector space need a refined estimation approach and further experimentation before we can definitively refute the hypothesis that multi-tiered word vector spaces can enable the identification of ranked neighbors as case recommendations with an optimal precision-recall balance that outperforms single-tier word vector spaces for the problem statement at hand.

6.2. Qualitative Evaluation

A qualitative human evaluation was conducted on a random sample of case titles from the held-out test set. The study involved 6 evaluators, 2 of whom are domain experts from GE Digital and 4 are AI/ML researchers from GE Research with experience in the power domain.

Each evaluator received the union of the set of nearest neighbors identified by each model, in randomized order, for each case title in the their respective evaluation sample set. The evaluators scored the relevance of each neighbor for a given test case title using three ordinal scores: “NO”, “PARTIAL”, and “YES”.

A common set of 5 test set case titles was provided to all evaluators to facilitate an Inter-Rater Agreement (IRA) evaluation. Each evaluator was required to evaluate neighbors for a minimum of 15 case titles, of which 5 were IRA case titles. Each evaluator was also given an optional set of 20 additional case titles to evaluate. We received evaluations for a total of 146 non-IRA case titles from the evaluators.

In addition to randomizing the ordering of nearest neighbors per test set case title, the ordering of the IRA case titles was randomized while ensuring the IRA titles appeared in the first 15 case titles provided to each evaluator. The randomization was done to prevent bias in the IRA and overall evaluation study.

6.2.1. Inter-Rater Agreement (IRA)

Fleiss’ kappa (Fleiss, 1971), a commonly used IRA metric, considers score labels or ratings as nominal or categorical in nature and, consequently, agreement/disagreement is viewed as binary as opposed to a continuous value. Since our evaluation involves scoring the nearest neighbors on an ordinal scale of three levels, Krippendorff’s alpha (Krippendorff, 2011) is more appropriate for computing IRA. Krippendorff’s alpha allows for agreement/disagreement to be non-binary when ratings are ordinal or have different severity of disagreement associated with them (Artstein, 2017).

The two IRA metrics computed pairwise across the evaluators in this study are shown in fig. 9. It provides a view of how the IRA is underestimated when Fleiss’ kappa is used for our evaluations, thus justifying the use of Krippendorff’s alpha. We observe an average Krippendorff’s alpha of 0.5546 for our evaluators on the IRA test cases.

6.2.2. Qualitative Evaluation Summary Discussion

We summarize the qualitative evaluation results by computing the Discounted Cumulative Gain (DCG) and precision

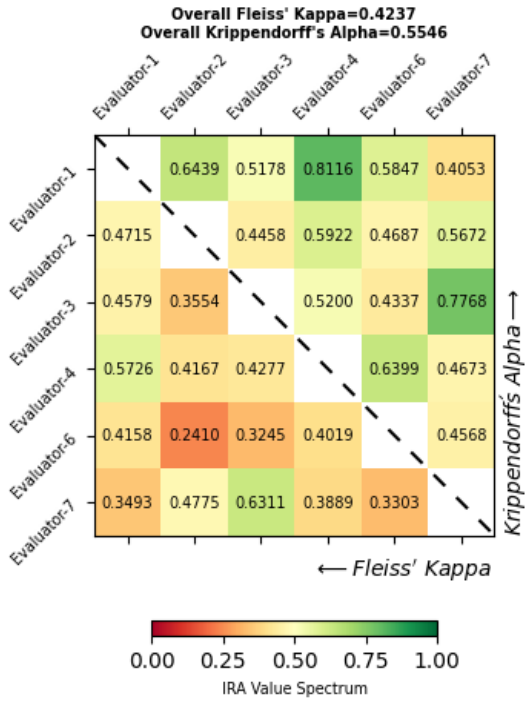


Figure 9. Pairwise Inter-Rater Agreement (IRA)

(Pr) for the sampled test set case titles for each model. The DCG and Pr are computed at every rank ($1 \leq r \leq 10$) in the rank ordered list of nearest neighbors that each model identifies by using the neighbors' corresponding evaluation scores assigned by the evaluators as ground truth relevance.

DCG provides a cumulative score for the ranked list of neighbors produced by each approach weighted by the neighbors' ranks. The DCG at a rank r ($DCG@r$) is calculated as follows:

$$DCG@r = \sum_{i=1}^{r} \frac{rel_i}{log_2(i+1)} \quad (1)$$

where, rel_i is the numerical relevance score of the neighbor identified at rank $1 \leq i \leq r$. DCG@r provides an understanding of the degree to which relevant neighbors are identified by each approach as we traverse the ranked list of neighbors. DCG increases when relevant results are ranked higher than irrelevant results and it decreases when relevant results are ranked lower than irrelevant results. The following numerical relevance scores are used for the three score labels from the qualitative evaluation to compute DCG: "NO"=0.0, "PARTIAL"=0.5, "YES"=1.0 .

The DCG@r plot in fig. 10 visualizes the identification of relevant results on average by the various models relative to each other as their ranked neighbor lists are traversed. The plot shows that the WVEC-1tier outperforms all other models. It is also the least complex of the four models and requires the least number of parameters, specifically only the

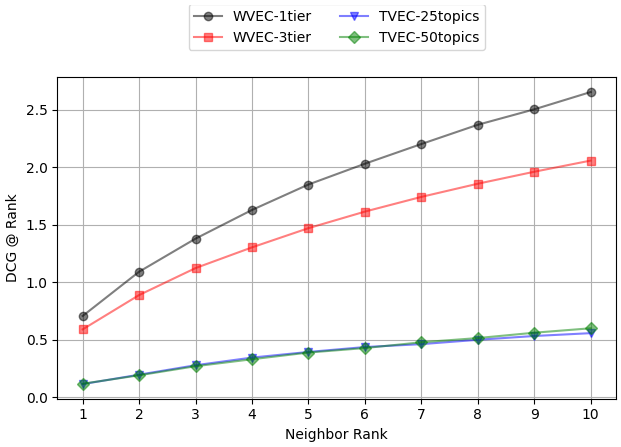


Figure 10. Mean Discounted Cumulative Gain (DCG) for the WVEC and TVEC Models

$\widetilde{IDF}$ filtering thresholds, making it an attractive model for the use case across domains.

The TVEC-$\mathcal{K}$topics models are comparatively the most complex of the models. They also under-perform both of the WVEC models. The DCG performance of TVEC-$\mathcal{K}$topics models show that vectorization of the case titles at noun-phrase granularity under-performs with irrelevant neighbors being identified in the topic vector space more often along with occasionally partially relevant neighbors.

The LDA model in TVEC infers topics as bags-of-noun-phrases that are optimized for discriminating case titles. The inability of TVEC models to identify relevant case recommendations indicates that the LDA models seem to optimize topics using uncommon noun-phrases while ignoring more common noun-phrases across case titles that would allow for effective case clustering. Further, topics are mixtures of terms which abstract out differences between case titles that contain in-topic but different terms, leading to unrelated case titles being clustered together.

Precision provides a measure of the proportion of relevant neighbors identified in the ranked list of neighbors. Pr@r quantifies the proportion of relevant neighbors identified at different rank (r) thresholds by the models. Since precision requires binary classification of the relevance of the neighbors, all neighbors scored as "PARTIAL" and "YES" relevance are combined to form the positive class. Pr@r is plotted in fig. 11 and shows the mean Pr@r for the different models.

The Pr@r plot shows the WVEC models identify relevant and partially relevant neighbors and rank them higher than the TVEC models. The low mean precision values for the TVEC models confirm the observation from the DCG plots that these models fail to identify relevant and partially relevant neighbors in the average case. Further, the trends ob-

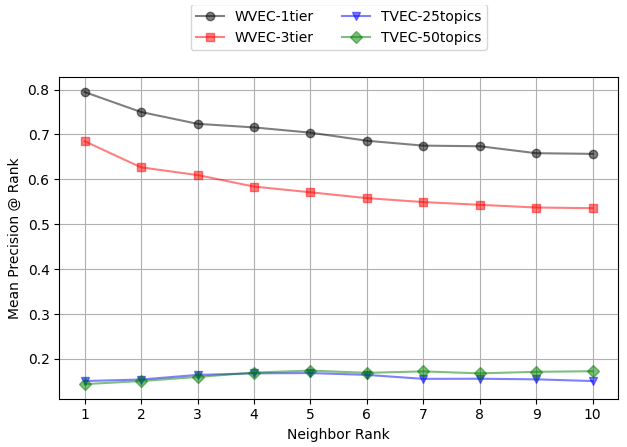


Figure 11. Mean Precision@Rank (Pr@r) for the WVEC and TVEC Models

served for mean Jaccard similarity (refer fig. 8) and Pr@r performance of the models are in close agreement and confirm the observations in section 6.1.

WVEC-3tier is better geared, in theory, to identify more relevant cases for an input case title since its neighbor search space is constrained to a small subset of related cases that belong to a cluster at a deeper tier. This is contingent on the choice of the $\widetilde{IDF}$ -based vocabulary segmentation boundaries specified as input parameters. Further evaluation of WVEC-3tier for a range of input parameter values as well using intelligent curve shape-based segmentation is planned for future work.

7. CONCLUSIONS

This work evaluated three different approaches for vectorizing maintenance case titles for kNN recommendation of cases relevant to a new input case title. Representation of case titles in simpler word vector spaces is shown to perform much better than a bag-of-noun-phrases topic vector space through quantitative and qualitative assessments. WVEC-1tier, the simplest of the three approaches we evaluated and with the fewest parameters, outperformed the other models. TVEC-$\mathcal{K}topics$ models, the most complex of the three approaches, perform poorly on this task.

Future Work

Future work to enhance these models includes the utilization of verbs, which should help them understand additional case context and component relationships for better case separation. In addition, the optimization of the $\widetilde{IDF}$ ranges for the WVEC-3tier model through further experimentation should help the algorithm better discriminate between cases more comprehensively and make it more competitive with WVEC-1tier. Finally, we intend to enable a full end-to-end pipeline that leverages the most effective vectorization approach to generate recommendations for enhancing the troubleshooting and maintenance workflows in order to achieve O&M cost savings through decreased time to case resolution and improved productivity.

ACKNOWLEDGMENT

The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0001290. 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 research team further acknowledges contributions from several other researchers and engineers who helped in annotating case titles to enable quantitative performance evaluation.

REFERENCES

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: Ordering Points to Identify the Clustering Structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data* (p. 49–60). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/304182.304187

Artstein, R. (2017). Inter-annotator Agreement. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 297–313). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-024-0881-2_11

Blei, D. M., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent Dirichlet Dllocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Edwards, B., Zatorsky, M., & Nayak, R. (2008). Clustering and Classification of Maintenance Logs Using Text Data Mining. In *Proceedings of the 7th Australasian Data Mining Conference - Volume 87* (p. 193–199). Australian Computer Society, Inc.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (p. 226–231). AAAI Press.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, *76*(5), 378-382. doi: 10.1037/h0031619

Hansen, B., Coleman, C., Zhang, Y., & Seale, M. (2020). Text Classification and Tagging of United States Army Ground Vehicle Fault Descriptions in Support of Data-Driven Prognostics. *Annual Conference of the PHM Society*, *12*(1), 8. doi: 10.36001/phmconf.2020.v12i1.1154

Hoffman, M. D., Blei, D. M., & Bach, F. (2010). Online Learning for Latent Dirichlet Allocation. In *Proceed-*

ings of the 23rd international conference on neural information processing systems (Vol. 1, p. 856–864). Red Hook, NY, USA: Curran Associates Inc.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. University of Pennsylvania. https://repository.upenn.edu/asc_papers/43

Nandyala, A. V., Lukens, S., Rathod, S., & Agarwal, P. (2021). Evaluating Word Representations in a Technical Language Processing Pipeline. *PHM Society European Conference*, *6*(1), 17. doi: 10.36001/phme.2021.v6i1.2894

Navinchandran, M., Sharp, M. E., Brundage, M. P., & Sexton, T. B. (2019). Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data. *Annual Conference of the PHM Society*, *11*(1). doi: 10.36001/phmconf.2019.v11i1.792

Ottermo, M. V., Håbrekke, S., Hauge, S., & Bodsberg, L. (2021). Technical Language Processing for Efficient Classification of Failure Events for Safety Critical Equipment. *PHM Society European Conference*, *6*(1), 9. doi: 10.36001/phme.2021.v6i1.2792

Pau, D., Tarquini, I., Iannitelli, M., & Allegorico, C. (2021). Algorithmically Exploiting the Knowledge Accumulated in Textual Domains for Technical Support. *PHM Society European Conference*, *6*(1), 12. doi: 10.36001/phme.2021.v6i1.2900

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408). Shanghai, China: Association for Computing Machinery. doi: 10.1145/2684822.2685324

Salo, E., McMillan, D., & Connor, R. (2019). Work Orders - Value from Structureless Text in the Era of Digitisation. In *SPE Offshore Europe Conference and Exhibition 2019, Aberdeen, UK*. Society of Petroleum Engineers. doi: 10.2118/195788-MS

Sexton, T., Brundage, M., Hodkiewicz, M., & Smoker, T. (2018). Benchmarking for Keyword Extraction Methodologies in Maintenance Work Orders. *2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA*, *10*(1). doi: 10.36001/phmconf.2018.v10i1.541

BIOGRAPHIES

Dr. Akshay Peshave is a Research Scientist - Machine Learning at GE Research. His research experience spans the breadth of machine learning with a focus on knowledge-based systems that address automated knowledge acquisition, representation, and reasoning. Akshay's current research includes maintenance action recommender systems for the ARPA-E GEMINA program, and explainability and interpretability of time-series forecasting models. His doctoral dissertation addressed unsupervised, domain-agnostic extraction of thematic phrases from single text artifacts. Akshay has 2 patents and 3 publications that address applications of machine learning methods to real-world problems at scale. He has been a key contributor to research projects at the Cognition, Robotics and Learning (CoRaL) Lab at the University of Maryland, Baltimore County led by Dr. Tim Oates that included research for the IBM Cognitive Horizons' accelerated cybersecurity program and the DARPA CREATE program.

Dr. Kareem S. Aggour is a Principal Scientist and the Knowledge Discovery Platform Leader at GE Research, where he informally leads a team performing research at the intersection of knowledge management and data management. His team develops innovative solutions to diverse challenges by applying techniques from knowledge representation and reasoning, natural language processing, and scalable data storage and computing. Kareem is leading the Work Order Automation effort on the ARPA-E GEMINA program, where he is partnering with colleagues from GE Research and GE Digital to build a technical language processing-infused case action recommendation pipeline. Kareem earned two B.S. degrees with honors from the University of Maryland, College Park, and earned both an M.S. and a Ph.D. from Rensselaer Polytechnic Institute while working full-time. Kareem has over 45 refereed publications and 24 issued patents.

Asma Ali is a Senior Staff Analytics Engineer with GE Digital where she is a technical team lead for the Analytics & Data Science team. Asma earned a Bachelors Degree in Biomedical Engineering from University of Connecticut and a Masters Degree in Mechanical Engineering from University of IL while working full-time. Asma has 15 years of industry experience focusing on software solution design, new product development and deploying process improvements, etc. Over the years, she worked closely with industrial domains like Power Generation, Oil & Gas, Mining & Metals, Aviation, Chemicals, Automotive, and more. Asma developed several dozen Digital Twins for Industrial Assets & Analytics for the APM product line. Currently, Asma is leading the efforts on Work Order Automation utilizing technical language processing for GE Digital and providing GE Research with Power domain-related industrial expertise.

Dr. Varish Mulwad is a Senior Scientist at GE Research with nearly 14 years of graduate school and industrial research experience in developing novel algorithms and production-ready solutions in the areas of information extraction and knowledge graph population from semi-structured and unstructured text. He has authored/co-authored 18 peer-reviewed publications, has 7 issued patents, and 875+ citations. He has experience in working across the spectrum of technology readiness level research programs, including deployed applications with $\sim \$10M+$ of estimated impact and has leadership experience managing & leading 3–4-member project teams. He received his Ph.D. and M.S. in Computer Science from the University of Maryland, Baltimore County. His complete CV can be found on http://varish.net.

Sharad Dixit is a Knowledge Software Engineer at GE Research, where he primarily focuses on designing and developing state-of-the-art solutions to various industrial challenges (aviation, power, healthcare, and renewables) in the areas of information extraction and natural language processing. He has authored/co-authored 8 publications and filed 1 patent. Sharad earned his M.S. in Computer Science from University of Maryland, Baltimore County, where he worked as a Graduate Research Assistant collaborating with the United States Naval Academy on the Delegated Access Control us-

ing Attribute-Based Encryption program sponsored by the Office of Naval Research.

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 the integration of AI-based PHM analytics in GE's industrial systems. He is the PI for the ARPA-E GEMINA program led by GE Research on 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, Abhinav 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. Abhinav 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 the International Journal of Prognostics and Health Management between 2011-2020. Abhinav actively participates in the organization of PHM Society conferences and various AI workshops on topics of Digital Twins and AI in Industrial applications.