Data-driven Approach to Equipment Taxonomy Classification

Kittipong Saetia[1], Sarah Lukens[2], Erik Pijcke[3], and Xiaohui Hu[4]

[1,2,3,4]*GE Digital, San Ramon, CA, 24013, USA*

kittipong.saetia@ge.com
sarah.lukens@ge.com
erik.pijcke@ge.com
xiaohui.hu@ge.com

ABSTRACT

A standardized taxonomy enables asset-intensive industrial organizations to systematically measure and track efficiency and performance of assets at different levels in an asset hierarchy. Having a well-structured taxonomy also allows companies to take advantage of emerging data-driven technologies, such as prognostics and health management (PHM), through enabling straightforward mapping of assets to analytical content specific to equipment commonalities, e.g., failure modes. However, the complexity and use of equipment taxonomy and coding structures in maintenance management systems vary widely for different organizations. This paper describes a data-driven approach for identifying equipment taxonomy from equipment records in maintenance management systems. The approach combines machine learning-based and rule-based methods into a hybrid man-in-the-loop workflow, which enables rapid and consistent mapping of equipment into a standard taxonomy. A case study is presented to demonstrate the performance and challenges of the proposed approach on equipment taxonomy classification.

1. INTRODUCTION

Business outcomes for asset-intensive industrial organizations depend on high performing and reliable equipment. There can be a massive amount of physical assets in a single industrial organization which can be organized or structured in a variety of ways such as by system, by function, or broken down into the smallest component. Having a categorization system in place to track all the assets in a registry by their function and physical characteristics helps to manage and keep track of assets. An equipment taxonomy is a structured way of classifying equipment into hierarchical groupings where levels are based on asset similarities. A taxonomy is a classification system for assets that facilitates indexing, grouping, saving, searching, and retrieving digital data (Fortin, 2018). Digitally, taxonomic classifications are typically captured as attributes to assets that are stored in an asset repository. Asset repositories are typically stored in central computerized systems for maintenance management processes such as the Enterprise Asset Management (EAM) or Computerized Maintenance Management System (CMMS) where the assets are linked to maintenance activities such as work planning, scheduling and execution (Distefano & Thomas, 2011).

Having a standard and consistent taxonomy in place at an industrial organization means there is one single data structure and methodology used for classifying assets across the organization. A standardized taxonomy enables a systematic way to use data for measuring and tracking asset efficiency and performance linked to different levels of the hierarchy. As a result, benefits include the ability to make comparisons across an organization as well as the ability to deploy "off-the-shelf" content at scale. For comparative analytics across units, sites, or plants to be meaningful, measures need to be related to comparable levels of the taxonomic hierarchy. For example, it does not make sense to compare the reliability and performance of a compressor train with a pressure transmitter. "Off-the-shelf" content which depends on specifics such as common failure modes which are context sensitive to the asset classification could vary from collections of standard lists and codes to templates for analyses such as Failure Mode and Effects Analysis (FMEA) to highly sophisticated machine learning models for prognostics and health monitoring trained for specific signals and failure patterns. In order to deploy any of this content at scale across an organization, a standard taxonomy is a pre-requisite for matching analytic content with appropriate industrial equipment.

While the benefits of a standard taxonomy seem straightforward, there are many issues which impede direct mapping from equipment registries in the EAM/CMMS to a standard. In extreme cases, an equipment taxonomy may not be used at all and elements characterizing an asset type can

only be inferred from various fields in the asset registry associated with individual assets. Other common poor practices include fragmented or incomplete entries when a taxonomic structure is used and the practice of grouping assets together under generic high levels in the taxonomy such as an entire system or unit. For example, all the assets in an air-cooling system may be classified as "air cooling system" instead of differentiating like-assets such as pumps, piping, tanks, heat exchangers, or cooling towers that may be part of the system. In cases where the taxonomies are meticulously implemented at a site or company, the taxonomy itself may differ between different sites and companies, which challenges both the execution of comparative analytics as well as the use of repeatable content and analytics deployed through a common system.

Information characterizing an asset can usually be found in the description fields for the equipment and/or functional location in the EAM/CMMS. Many challenges regarding missing or inconsistent taxonomic classifications can be addressed through utilizing the unstructured description of the asset in the asset registry. Methods from natural language processing (NLP) can be utilized as a means to structure the unstructured text. There have been ongoing efforts to utilize methods from NLP on maintenance work order data for applications such as characterizing failure events from field maintenance data for reliability analytics (Sexton & Brundage, 2019) (Sexton, Brundage, Hoffman, and Morris, 2017) (Sexton, Hodkiewicz, Brundage, and Smoker, 2018) (Hodkiewicz, Batsioudis, Radomiljac, and Ho, 2017) (Hodkiewicz & Ho, 2016) (Arif-Uz-Zaman, Cholette, Li, and Karim, 2016) (Lukens, Naik, Doan, Hu, and Abado, 2017) (Lukens & Markham, 2018) (Lukens, Naik, Markham, and Laplante, 2019) (Lukens & Markham, 2018).

There are many similarities in the use of NLP to characterize maintenance work order data with characterizing equipment taxonomy that differ from NLP algorithms applied to consumer world applications. As a result, available NLP packages which are developed for consumer applications often produce inadequate results when applied to industrial text data. Industrial text often contains or refers to implied domain knowledge which does not exist explicitly in the text itself such as jargon, acronyms/abbreviations, or industry specific context.

This paper reviews the importance of a standard equipment taxonomy, summarizes challenges with imposing a standard, and proposes an approach for mapping equipment to a standard taxonomy using the equipment description fields. The rest of this article is organized as followed. Section 2 summarizes existing standards for equipment taxonomy as well as data quality challenges common in equipment registries. Section 3 describes the high-level workflow that we have adopted based on the challenges in the data. Section 4 presents a case study illustrating the methodology. The paper ends with concluding discussions and suggests future research directions.

2. BACKGROUND

2.1. Equipment taxonomy and existing standards

An equipment taxonomy is a systematic classification of equipment into a hierarchical grouping where levels are based on the complexity of equipment characterization (ISO, 2004). For example, "rotating equipment" is a high-level characterization which includes pumps, turbines and compressors but has similarities that differentiate this grouping from other high-level groupings such as instrumentation which includes sensors, transmitters, etc., or fixed equipment which includes pressure vessels, heat exchangers, tanks, etc. Pumps, which are a class of similar rotating equipment, can further be split into centrifugal pumps, piston pumps, peristaltic pumps, screw pumps, etc. Pumps all have certain commonalities, but there are differences between the different types of pumps. And each type of pump has its own set of components, failure modes, etc. which may be similar or different. For example, both centrifugal pumps and reciprocating pumps are used for pumping liquid and typically experience "seal leaks" as the most common failure mode. However, the mechanism which drives these pumps is different – reciprocating pumps push liquid through positive displacement forced by a piston while centrifugal pumps continuously pump through kinetic energy of its impeller. It is important to be able to both differentiate these as two separate equipment groups as well as group both together depending on the use case, which is possible when an equipment taxonomy is in place.

The equipment taxonomy differs from the functional location hierarchy, which is another important hierarchical grouping of industrial assets. The functional location hierarchy describes the physical relationship and hierarchy of assets, such as from site to unit to area, and also plays a distinct and important role. It is important to distinguish the difference between the serialized asset and its functional location (Distefano & Thomas, 2011). While the taxonomy will always stay fixed for a serialized asset, the functional location may change. For example, if the asset is a repairable spare it may be removed from service, sent out for re-build, and possibly re-installed in a different functional location. Asset characterizations at this level are important for being able to identify issues with repairable spares. If failures or performance is not tracked against both the serialized asset and the lowest level of the functional location, it is extremely difficult to determine if a recurring issue is associated with a particular location or a poorly-rebuilt asset. Benefits of a functional location hierarchy include ability to locate assets in a database when creating work orders, allows roll up of performance metrics by physical and functional location, and

facilitates organizing and grouping of assets within a location.

Standards and best practices do exist for equipment taxonomy but vary across different industries and verticals. One well used standard in the oil and gas industry is ISO-14224 (ISO, 2004). ISO 14224 in general provides guidelines for collecting maintenance and reliability data in a standard format. ISO 14224 defines taxonomy with respect to recommendations about levels and boundaries and provides a suggested standard taxonomy for common assets in upstream oil and gas. While developed for offshore structures in oil and gas industries, other companies both inside and outside the petroleum, petrochemical, and natural gas industries use ISO-14224 as a standard for formatting and coding equipment taxonomies (Distefano & Thomas, 2011). The KKS Identification System for Power Plants is a classification system for power plants (Konigstein, 2007). The KKS system includes codes identifying plant equipment and its operation ranging from building, structures, and equipment levels which can be used for developing databases for plant maintenance and management. The KKS system is the most widely used in European countries. The commercial aircraft industry uses the concept of "ATA Chapter", which are documented in the Joint Aircraft System/Component (JASC) Code Table (JASC, 2002). ATA chapters categorize the various systems that are on a plane such as "Communications", "Equipment/Furnishings" and "Pneumatics" and are used in the aviation industry as a standardized system to code the different technical features of an aircraft.

Currently to date there is no one standard taxonomy that captures all industrial use cases. Due to the wide variety of industrial verticals, equipment, and applications, it is likely that if one "standard" is ever adopted, it would look more like a standard process for managing a centrally stored taxonomy than a static list. The details of a single standard taxonomy and how to manage such a list is an open research topic, important for all of the reasons described in this paper, but beyond the scope of this work. This work assumes that a standard taxonomy which is sufficiently-good to capture the given data for the desired use case exists. The case study presented in this work specifically uses a home-grown taxonomy for equipment that originated from ISO-14224 but has evolved as data from other verticals have been included as the "standard taxonomy". This taxonomy has proven in practice to be very effective for consistently describing equipment across hundreds of industrial companies in the process, power generation and mining industries. The equipment levels used are category, class, and type, respectively.

2.2. The relationship between analytics and equipment taxonomy

ISO 14224 (ISO, 2004) observes that the level of equipment taxonomy is directly related to the types of analyses possible, and in order to conduct a certain type of analysis, there must be information at the required level. Reliability and maintenance data need to be related to a certain level within the taxonomy hierarchy in order to be comparable in a meaningful way. Determining the appropriate level of the taxonomy for an analysis depends on what the challenge/business problem is that the analytic is trying to solve. For example, analyses that link maintenance with operations such as availability, production stoppages, or quality may be more meaningful at higher levels, while analytics that are dependent on failure mode patterns and particular sensor measurements are very specific to the type of equipment. Different analytic characterizations are reviewed below.

Metric evaluation for benchmarking and comparative analytics. Metrics or key performance indicators (KPI's) are measurable values which can be used both to evaluate the effectiveness of a company to meet its goals and to identify opportunities by comparing like-assets. For example, if one particular asset is failing three times as much as similar assets, comparative analytics offer a way to identify and measure the opportunity. Depending on the use case and which metrics are used, it is important to be able to group like-assets under any comparable level of both the equipment taxonomy and the functional location hierarchy.

FMEA or asset strategy template. FMEA or strategy templates are templates which list possible failure modes and for each failure mode, quantifies the probability and consequence of each risk and lists the possible mitigating actions. FMEA or Strategy templates are useful as a starting point for creating maintenance strategies (instead of starting from scratch). Linking lists of possible risks and suggested mitigating actions must be done at the equipment type level.

Downtime and availability analysis. Formally, availability is the probability that an asset can perform its intended function satisfactorily when needed (Gulati, 2013). Availability is a key input measure for estimating performance efficiency measured by Overall Equipment Effectiveness (OEE) along with quality and performance as well as being central to estimating production losses. Some critical and expensive assets and systems can cost a facility "money-by-the-minute" when they are unavailable. In many cases, measures of availability may make more sense when evaluated at the system or unit level because the system or process level is the level that impacts process or plant efficiency. In other cases, such as identifying improvement opportunities through comparative analytics, availability may make more sense at the asset type level, at the failure mode

level, or at different levels of the functional location hierarchy.

Weibull/Reliability distribution analysis. In Weibull analysis, statistical distributions are fit to time-to-failure (TTF) data samples in order to make predictions about the expected failure times for an equipment population. Common statistical distributions used to fit TTF data include the Weibull distribution, the Exponential distribution, and the Lognormal distribution. Different failure modes have different time-scales and failure properties, so each failure mode should be modeled and characterized by its own statistical distribution parameters. For example, some failure modes may tend to occur earlier in the lifetime of an asset, such as failure modes due to improper installation or manufacturing defects. Other failure modes may be related to the wear-out or degradation of an asset and are observed later in life. For statistical distribution fitting, it is essential to have failure mode information to fit a statistical distribution (Abernathy, 2004), (Meeker 1998).

Reliability Growth analysis. While statistical distribution analysis is useful for TTF predictions for non-repairable assets at the failure mode level, there is also a class of statistical models for repairable systems based around Non-Homogeneous Poisson Processes (NHPP) referred to as Reliability Growth models in the reliability community (Abernathy, 2004), (Meeker & Escobar, 1998). Reliability Growth models are used when the time between one failure to the next is dependent on the time between the previous failures for a repairable asset or system. The most common Reliability Growth model is the Crow-AMSAA model, which is an NHPP model with the recurrence rate described by a power-model. While there is rich statistical theory behind the Crow-AMSAA model, in practice, parameters can be simply estimated and the model is very robust for applications where understanding recurrent events is of interest. The Crow-AMSAA model is useful for estimating next event times for many different event types of interest ranging from failures due to a specific failure mode to general corrective maintenance events. For this reason, depending on the application of interest, the Crow-AMSAA model can be used at different levels of the taxonomy. However, for expected failure predictions the model should be at the asset type level.

Analytic template/blueprint for PHM analytics. Typically, diagnostic or prognostic models are built or trained from readings from condition-based monitoring (CBM) which give measures of system health which can be used to report the current state of an asset or for determining performance life remaining in an asset. Different CBM technologies may measure different attributes such as flow rate, temperature, pressure, vibration, etc. which give indications of system health. For instance, a change in differential pressure on an air intake filter in a gas turbine could be due to the filter fouling, or increased bearing temperatures on a pump could be due to degraded lube oil impairing cooling of the bearings. Models which can capture these measures of system health are very specific to the asset, the CBM readings, and possible failure modes. In order to deploy such a model at scale there needs to be some sort of templatized way to consistently match the analytical models to the appropriate assets and signals.

The different analytical categories discussed above and the required levels needed in a standard taxonomy and in the functional location hierarchy to deploy these analytics are shown in Table 1. For simplicity, the functional location hierarchy is abstracted to two levels; a site or plant level and a lower meaningful level describing a functional process within a site or plant which can be a unit, an area within a unit, or a line. In practice the functional location hierarchy is a key hierarchy by itself that needs to be well defined, be consistent across different sites and should standardized in the context of the equipment taxonomy, but those levels of granularity are not required for the information presented in the table. The levels of the equipment taxonomy are expressed by Category/Class/Type. While there is a component/maintainable item level below the equipment type, in practice analytics as content would be applied to the asset type. It is important to remark that missing from this table (Table 1) is the actual asset itself. Specifically, what is missing is the relationship or mapping between the serialized asset to the lowest level of the functional location hierarchy, and determining whether an analytic should be mapped to the serialized asset or to the functional location is another decision to make based on the application and desired output.

2.3. Data quality challenges in taxonomy

There are data quality challenges when mapping a standard taxonomy to general equipment registries from the CMMS/EAM across different sites or companies that can make the process not straightforward. One data quality challenge is that sometimes an industrial company will not use an equipment taxonomy at all. The information about the asset in the equipment registry may not be based on taxonomic coding and instead relies on unstructured descriptions, which in turn may contain misspellings, abbreviations, or reflects general poor data entry practices. A few examples of cases where the equipment type is not coded, but there is information in the description field that could possibly is shown in Table 2.

Table 1. Mapping different types of analytics that could be developed and deployed as content to the required levels of the taxonomy and/or functional location hierarchy needed.

Analytics as content	Functional location hierarchy		Taxonomy			Notes
	Site/Plant level	Unit/area or line level	Equipment category	Equipment class	Equipment type	
Benchmarking and comparative analytics for opportunity identification	X	X	X	X	X	Ideally want all levels to match to enable drill down analysis
FMEA or asset strategy template					X	Requires common failure modes and actions
Analytic template/ blueprint for PHM analytics					X	Requires common failure modes and sensor readings
Downtime and Availability analysis	X	X	X	X	X	Level depends on the application
Weibull/Reliability distribution analysis					X	Requires similar failure modes
Reliability growth analysis	X	X	X	X	X	Level depends on the application
Impact of failure or maintenance on safety	X	X				Ref: ISO-14224
Impact of failure or maintenance on operations	X	X				Ref: ISO-14224

Table 2. Examples of equipment descriptions in the case where the taxonomy is unavailable (no codes)

Equipment Short Description	Equipment type (code)	Notes
Fork Lift, Manufacturer name	NULL	Fork Lift
GAS TURBINE	NULL	Gas Turbine
VALVULA DE SEGURIDAD	NULL	Safety Valve
Feedwater pump	NULL	Unable to determine at a level lower than "Pump"
ABC123-TO-BE-EDITED	NULL	Unable to determine
Unknown	NULL	Unable to determine

The examples in Table 2 illustrate that while often the required information is directly found in the text, often it is not possible (such as "Unknown", or "to be edited"). Further, while information about the equipment taxonomy may be present, it may not be enough to populate the full taxonomy such as "feedwater pump" or "valve". It is possible to determine that a "feedwater pump" is a pump, but not what type of pump. It is very common for the equipment description to contain information about the service or location of an asset.

When the taxonomy codes are available, often the structure is fragmented or incomplete. For instance, assets may be coded at too high of a level such as "PIPING", which may contain tens of thousands of pipes, hoses, and supports. This is a barrier to enabling the analyst to roll up and measure reliability, asset performance, or cost information. For example, Sikorska, Hammond, and Kelly (2008) were faced with the challenge of an equipment taxonomy which was not constructed to assign a failure to the correct level (In this case, it was too high level, so in the default structure, could only assign blocked lube oil filter to the whole lube oil system). Sikorska et al. (2008) solved this by creating a taxonomy based on ISO 14224 for the data, and stored mapping rules linking the standard taxonomy to the CMMS, and to the database with reliability content in order to link content to the appropriate equipment.

2.4. Natural language processing applied to industrial data

NLP has been increasing in popularity as tool for characterizing maintenance work events from maintenance data in industrial applications. Many key challenges in maintenance data which make straightforward use of analytics are also applicable to equipment taxonomy. Different data cleansing approaches and challenges for maintenance data are extensively reviewed by Hodkiewicz and Ho (2016), which are summarized here in the context of equipment taxonomy.

Rule-based systems can be used to quickly characterize a high volume of text sufficiently well, but managing a rule set can grow into increasing complexity and size as data grows. Rule-based systems have been used successfully by Sexton et al. (2018), and Hodkiewicz and Ho (2016). Syntactic processing such as keyword spotting is a great way to manage and generalize rule sets by developing scalable ways to manage keyword lists, but are challenged by ambiguities that may occur. For example, the word "motor" can describe a motor, or it can be part of something such as "motor-controlled" or "motor control center". The word "pressure" alone does not have meaning as an equipment type, but becomes essential in the context of "pressure transmitter", "pressure vessel", "pressure relief valve", etc.

Machine learning classification is very powerful for making predictions based on training data, and has the ability to scale with increasing data sizes or complexities. The challenge with classification models is that there needs to be a labeled training data set, and the original challenge is lack of labeled data (Arif-Uz-Zaman, Cholette, Ma, and Karim, 2017) (He, 2016) (Sexton et al., 2017). Assuming there is a sufficient volume of labeled data, the first step to building a classifier is the process of converting the words into input to machine learning classifiers, or vectorization. The most common way for representing text is through a "bag-of-words" (BOW) model (Cambria & White, 2014), where words or tokens in the text are features. However, there are many ambiguous words in equipment descriptions that could confuse such a model, for example if the description contains a higher-level system (such as instrumentation on a turbine or compressor), or ambiguous words that can have several meanings. With recent advances in recurrent neural networks (RNN) and word embedding algorithms such as word2vec (Dos Santos & Gatti, 2014) (Goldberg, 2016), there exists the potential to handle ambiguous words based on context in the description. However, new challenges emerge because a massive amount of labeled data is required for these models to perform adequately.

Class imbalance is a characterization of equipment taxonomy data by its very nature. There may be one or two gas turbines at a powerplant, but thousands of valves, instruments, circuit breakers and piping. Out-of-the-box machine learning classifiers are challenged by class imbalance which must be addressed for adequate model performance. Not only does the data feature class imbalance, but the example also shows how two equipment may not be equal. There should be emphasis on correctly labeling the gas turbine, but correctly capturing two pipes or breakers at the lowest level of the taxonomy may not be as important.

A challenge for both training models and for enabling comparative analytics is consistency between different sites, units, or companies. The information needed to accurately characterize taxonomy may not be found in the description field, but found in another field somewhere else in the CMMS/EAM data. A couple of examples are shown in Table 3. The first two rows show the case where identical description fields may have completely different codes, while the last two rows show how the description field alone may not contain the level of granularity to infer equipment type. While this is okay in general since the equipment is coded and present in the registry, if this labeled data is used to train a model to apply to new data, challenges emerge. If the fields are combined in the training, the prediction model may be overfit, but if the fields are not combined, key information may be lost. For these reasons, different approaches should be used depending on the end goal. Two distinct end goals for taxonomy classification are: (1) predicting or filling in standard taxonomy values based on the data descriptions, or (2) standardizing equipment taxonomy in the cases where codes are inconsistent using the unstructured descriptions as well as provided codes.

Table 3. Examples when the description fields alone do not capture the appropriate information needed to classify equipment taxonomy.

Equipment description	Equipment Type (Code)
Regeneration pump	Reciprocating pump
Regeneration pump	Centrifugal pump
HX-1	Heat Exchanger - Plate
HX-2	Heat Exchanger – Shell and Tube

The "no free lunch" theorem applies to characterizing equipment taxonomy. We propose a hybrid approach which combines machine learning classification algorithms and rules-based methods in a workflow with a man-in-the-loop, described in the next section. The machine learning models are data-driven and will scale as new data is added, but require the data to be cleaned and labeled. The rule-based methods are knowledge-driven and incorporate domain-expert knowledge to make inferences on the data in a scalable way. The scope of the workflow focuses specifically on the first end goal: to predict standard taxonomy values using the description fields where none exist.

3. METHODS

We describe the end-to-end workflow and considerations for mapping equipment data to a standard taxonomy based on the data characterizations from the unstructured description fields in the asset registry. In this case it is assumed that there is already equipment data mapped to standard for sites in the same vertical which can be used to train a model. Due to having training data, at a high level, the workflow is based on a standard machine-learning training-prediction workflow, where labeled data is used to train a classification model and the model is then used to make predictions on unlabeled data. However, due to the many challenges in the equipment data, many considerations need to be addressed and modifications developed in order to execute the workflow. The high-level workflow is shown in Figure 1 below.

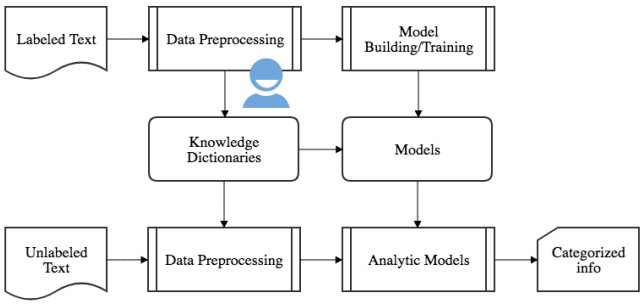

Figure 1. High level standard workflow for training classification models on labeled text, and using the models to make predictions on unlabeled text.

In the standard workflow, labeled data containing equipment descriptions with labeled codes are used to train a classification model using supervised machine learning. The first step for developing the model is data pre-processing. For classification models, this involves not only traditional text cleaning steps such as removing special characters but also word replacement using knowledge dictionaries for feature selection/reduction. Specific knowledge dictionaries capturing equipment words and phrases need to be developed, and a man-in-the-loop component is introduced to develop custom dictionaries for equipment words. The need for knowledge dictionaries arises from the fact that domain-specific acronyms/abbreviations and misspelling are common in equipment data, e.g., xmtr is used in place of transmitter; condensor and condensors are used to describe condenser. Once pre-processing is conducted using a combination of rules and knowledge dictionaries on the labeled data, classification models are trained and selected. In developing these models, challenges with quality of labeled data emerge and are addressed through introducing another man-in-the-loop component.

For unlabeled (new/unseen) data, the expected pathway is to run the data preprocessing pipeline that was developed from the labeled data, and then runs the trained model on the cleaned text to make predictions. In this case, potential incompatibilities of the unlabeled data with the trained model emerge (a symptom of lack of sufficient labeled data). The workflow is modified to determine which predictions are sufficient, and which predictions need review.

3.1. Model training on labeled data

3.1.1. Data pre-processing

Data-preprocessing is essential for any NLP model, but is particularly important for unstructured text in industrial applications because prevalent spelling errors, abbreviations and ambiguities typically lead to many distinct words which have the same meaning. Having different words describing the same word adds noise to the data, and this leads to classification models with poor quality. Data pre-processing includes both rule-based and dictionary-based steps for cleaning text, feature selection, and keyword extraction. Rule-based methods can be blanket applied, while dictionary-based require managing and maintaining a dictionary. Text pre-processing can also be classified as out-of-the-box, which can apply to any industrial dataset, or specialty which is specific to a dataset or application. These 4 combinations of text pre-processing steps are summarized in Table 4. We utilize the following sequence of text cleanup steps for equipment text data: 1) rule-based specialty, 2) rule-based out-of-the-box, 3) dictionary-based out-of-the-box, 4) dictionary-based specialty. This sequence was found to work well with the equipment data.

Table 4. Examples for each of the 4 different combinations of rule-based versus dictionary-based text cleaning against generic out-of-the-box and specialty.

	Out-of-the-Box	Specialty
Rule-based	• Lowercase • Remove special characters/numbers	• Special case rules such as remove text in (), remove equipment ID
Dictionary-based	• Remove stop words such as “are, the, is” • Expand contractions such as “isn’t” to ”is not”.	• Removal or replacement of text based on a specific case, such as site-specific jargon or abbreviations

Creating dictionaries based on the use case can be used for feature engineering, and for keyword extraction. We use the human tagging process developed at the National Institute of Standards and Technology (NIST) (Sexton et al., 2017)

(Sexton et al., 2018) (Sexton, T., Moccozet S., Brundage, M. P., Madhusudanan, N., Hastings, E., & Bones, L). The process catches similar terms, as well as merging important N-grams to 1-grams (such as "circuit breaker", "motor control center", and "relief valve"). Mapping instances of "motor control center" to "motor_control_center" removes ambiguities around those instances of the word "motor". The tagging process also identifies stop-words to remove to contribute to feature reduction in the cleaned text. We manage the dictionaries in a central database so changes are applied immediately in the pipeline.

3.1.2. Inconsistent label handling

One of the main challenges in training classification models is due to the inconsistent labels of training data as described in Section 2.3. Modifications to the work process include adding a data quality check step to automatically flag problematic data points with inconsistent labels, and a subject-matter expert (SME) to review the labels. The work process for handling inconsistent labels is shown in Figure 2. After the data goes through the pre-processing stage, it goes through a data quality check. The data quality check is composed of rules such as if the descriptions are identical but the labels are different, create a flag to review. Once the flagged entries are reviewed, the training data is preprocessed and consistent and ready for model training.

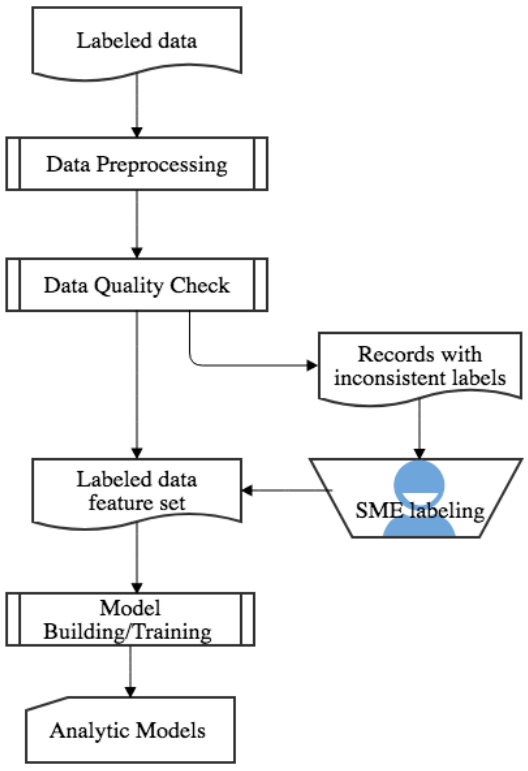

Figure 2. Inconsistency check for labeled data.

3.1.3. Classification engine

We use a supervised machine learning approach to build and train classification models to predict equipment code from equipment text descriptions. We explored two families of models; classification models using bag-of-words based on the sklearn package (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay, 2011), and RNN models using word embeddings based on keras/TensorFlow (Chollet, 2015). For the bag-of-words model, support vector machine (SVM) and random forest (RF) algorithms were considered. The workflow was built to incorporate flexibility for different classifiers for model selection.

3.2. Unlabeled data modifications

Once the model was trained on the labeled data, the next step is to run the pipeline on the unlabeled data. However, due to challenges discussed in Section 2.3, depending on the size, equipment, and vocabulary used in the labeled training dataset, it may be inappropriate to predict all of labels for the unlabeled data from the training data. Classification models always return a prediction, and while many are valid, many can also be invalid. The workflow aims to utilize the valid predictions, necessitating the development of an approach which automates the determination of which predictions are trustworthy and which are not based on the prediction probability and through a compatibility score.

The compatibility score developed here is the ratio comparing the number of features (words) in a new description with the vocabulary from the training set. For example, for a description 5 words long, if 4 of the 5 words are features in the training set, the compatibility score would be 4/5 = 0.8. The compatibility score is a quick way to determine if the equipment described is relevant to the training set or if it not.

The workflow for the compatibility check is illustrated in Figure 3. Once the unlabeled data is pre-processed, the compatibility score is calculated for each cleaned description. For compatible descriptions with a satisfactory prediction probability from the trained model, a prediction result is produced. For records with low compatibility scores (features that are not captured in the training data), an active learning approach is implemented where the SME labels the feature set to be fed back to the trained model and the process resumes.

4. CASE STUDY

In the case study, we walk through the different steps of the workflow. The case study illustrates how the process uses labeled data from with 73,000 assets and 110 different equipment types to train a classifier and develop preliminary word dictionaries, which are then applied to unlabeled data at

4 new sites with 42,000 assets. To protect proprietary information, all variables have been anonymized and numerical quantities such as counts in classes have been mixed. Enough data is used in the case study to train a model, but not enough data to capture all of the features in the unlabeled data.

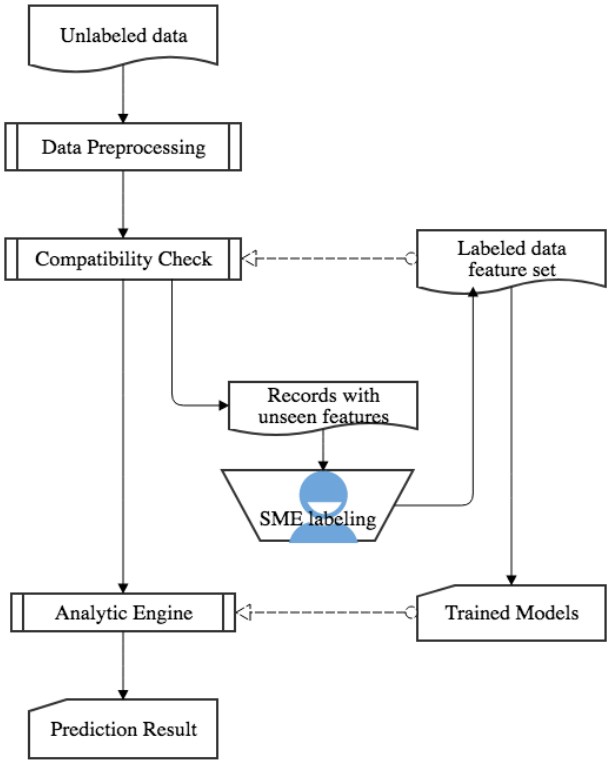

Figure 3. Compatibility check for unlabeled data.

4.1. Pre-processing

After the preprocessing workflow was executed, the number of unique words (features) in the training data was reduced from about 3,200 features to 2,400 features. Examples of how the text cleaning reduces features is shown in Table 5 below. For instance, alpha-numeric characters such as "1ABC-YZ-23" are removed and N-grams such as "RELIEF VLV" and "Relief Valve" are mapped to "relief_valve" after the dictionary-based preprocessing step based on human tagging.

4.2. Model training

The labeled dataset has 73,000 records with 110 classes of unique equipment types. Model performances for support vector machine (SVM) and random forest (RF) are summarized in Table 6. We compare the accuracy and F1 score from a test set with 0.7-0.3 train-test split. In this setup, 70% of the labeled data was used to train models, and 30% of the data was held out to test performance of the trained models. The SVM model had higher accuracy and F1 for the test set, but these values are not impressive at ~70%. Several factors contributed to the low performance of the test set including class imbalance. 12% of the class labels contained 80% of the data in this dataset. Further exploration revealed that class imbalance was not the most significant driver of low accuracy, it was inconsistent labels.

Table 5. Equipment descriptions before and after text cleaning steps, notably application of the corpus specific dictionary developed using human tagging.

Equipment type	Raw description	Cleaned description
Relief valve	1ABC-YZ-23 FILTER PRESS AIR RECEIVER RELIEF VLV	filter press air receiver relief_valve
Relief valve	HRSG-1 HP Drum Pressure Relief Valve, PSV 1234	hp drum pressure relief_valve psv
Temperature indicator	HP STEAM TEMPREATURE INDICATOR	hp steam temperature_indicator

Table 6. Performances of SVM and RF models on labeled data.

	Support Vector Machine		Random Forest	
	Accuracy	F1	Accuracy	F1
Train set	85.1	0.849	99.1	0.991
Test set	71.0	0.703	68.6	0.682

We explored using word embedding algorithms with RNN's using TensorFlow, but did not have better results than traditional machine learning algorithms, e.g., SVM and RF. Several reasons may have contributed to this result. One, neural network models require significantly larger training datasets than the dataset we had. Second, inconsistent labels in the data are a high influencer of model performance. The first level of performance gains was to focus our workflow on improving the labels of the training dataset.

4.3. Improving inconsistent labels

We utilized the main-in-the-loop workflow in Figure 2 from to clean up some inconsistent labels. A subject-matter expert (SME) manually reviewed texts in the equipment description field and modified the equipment class label field if the label was incorrect. One SME spent about 3 hours reviewing inconsistent labels and, in that time, reviewed and labeled 670 samples. A preliminary result of model performance from a test set using the 670 manually reviewed samples is shown in Table 7. The test-set accuracy of data with reviewed labels is

significantly higher than the original labeled data for both SVM and RF. This result is not surprising because machine learning models do not generalize well if the models were trained using poor quality data with unreliable labels. To further improve the generalizability of our models, additional effort is required to review and improve the labels of the training dataset.

Table 7. Comparison of performances of SVM and RF models trained on the 670 manually reviewed samples before and after the labels were reviewed by SME by the consistency check.

	SVM		Random Forest	
	Accuracy	F1	Accuracy	F1
Original labels	55.0	0.53	53.0	0.518
Reviewed labels	87.5	0.872	83.9	0.833

4.4. Predictions on unlabeled data

We applied the workflow described in Figure 3 to make predictions on unlabeled data containing 42,000 records. The SVM model described in Section 4.2 was used to classify the unlabeled data. The labeled and the unlabeled data each contained about ~2,000 unique words (features) after text preprocessing step. A compatibility check comparing the training corpus with the unlabeled corpus revealed that only about 1,000 of these features were in common. Several factors contributed to the low compatibility including some equipment descriptions were found to be in Spanish, and the labeled data insufficiently covered the various equipment in the unlabeled data. For example, the labeled data was not exhaustive enough to capture all of the equipment on the standard taxonomy, and some equipment types were in the unlabeled data that were not in the training data. These equipment were filtered out by the compatibility score.

For records with low compatibility (<0.8), a manual review by an SME was performed to label features such as stop-words to reduce feature size and improve training data. For unlabeled data with high compatibility, the classification results of the trained model were candidate labels. Predictions with high predicted probability scores (>0.7) were auto-labeled with confidence while others with low scores also were candidates to be reviewed by an SME.

Table 8 shows some examples of predicted labels and scores by the SVM model on unlabeled data. The first two examples have a high compatibility of 1 where all the words had been seen in the training data. They also have high prediction score, so we can confidently use these results. The third example has a low prediction score and needs to be reviewed by an SME. The fourth example has low compatibility and prediction score; thus, a manual review is needed.

From the unlabeled dataset, 28% of the equipment descriptions were auto-labeled by the trained model using the compatibility and prediction score criteria. To measure performance of the compatibility check process (Figure 3), 600 unlabeled descriptions with the lowest compatibility scores were reviewed and labeled by an SME. The classification model was retrained with the additional 600 newly labeled samples. The number of equipment descriptions auto-labeled by the compatibility and prediction score criteria rose to 33%. While results show promise for the man-in-the-loop workflow, the low numbers (only about 1/3 of the unlabeled data got auto-labeled by the trained model) signify need for additional work, e.g., gathering larger labeled data with a diverse range of equipment, before such a process can be deployed at scale.

Table 8. Examples of predictions and scores by SVM model on unlabeled data.

#	Cleaned equipment description	Compatibility	Predict label	Predict score
1	pump purge check valve	1.0	Check Valve	0.80
2	piping	1.0	Piping	0.81
3	air receiver	1.0	Storage Tank	0.19
4	weatherproof addressable breack glass	0.25	Valve	0.25

5. CONCLUSION

The workflow proposed in this paper discusses how NLP for industrial data can be used for standardization of equipment taxonomy using equipment descriptions. The workflow can be used not only for identifying missing fields, it can also be used to make existing labeled data more consistent, and can be used in companies with multiple sites in order to enable comparative analytics. Considerations, challenges, and requirements needed to implement and deploy the workflow include man-in-the-loop to review data in a prioritized way (active learning approach), and data quality challenges with both labeled and unlabeled data.

The learning model approach with a feedback loop is an advantage of this process. Another strength is utilizing the best of both worlds in multiple cases. One case is taking the strengths of a machine learning model, which is ability to scale and capture complex patterns in the training data, with the strengths of a rules-based model which does not require labeled data, only expert knowledge. However, due to the

low percentage of auto-mapped records in the case study, early on in the process there should probably be more emphasis on developing and using the appropriate rule-based model to get the numbers higher before focusing on the data-driven models. Another example of strengths and weaknesses is using the strengths of a person, which is looking at a description and knowing what is, and the strengths of a computer, which is to consistently execute extremely quickly.

Weaknesses of this approach can be viewed into two broad categories: technical weaknesses and functional weaknesses. One challenge from technical perspective is model selection for classification models. For the smaller training data such as in this study, bag-of-words with SVM was sufficient, but as the training data grows in size and complexity, when should the modelling approach change? Measuring the accuracy on the unlabeled data is another challenge. Technically, there are statistical methods that could be applied to infer the accuracy by sampling the unlabeled data and reviewing the predictions. But functionally this approach becomes challenging. Because the taxonomy is a hierarchy, if a model correctly captures the class but does not get the type, is it correct? Is it correct when the type is not in the description, but incorrect when that level of information is in the description? For example, if one description says "valve", and the model predicts the asset is a generic valve in the valve class, and another description says "ball valve" and the model also predicts the asset is a generic valve in the valve class, but "Ball Valve" is an asset type on the standard list, is that incorrect? How do you treat these cases when evaluating accuracy of the predictions? In practice, this will depend on the purpose of why the taxonomy is getting mapped, which may vary from case to case.

A key assumption made in this work was that a satisfactory standard taxonomy already existed and was available. As mentioned in Section 2.1, the details of a standard taxonomy and how to manage such a list that is applicable across industries is an open research topic. One lesson we learned in this study is that there is need to develop approaches to manage and grow standard taxonomies based on what is observed in the data as well as approaches to synchronize breathing standard taxonomies with asset registries. Such work is an important research topic that would have usefulness across all industries.

ACKNOWLEDGEMENT

The authors acknowledge Manjish Naik, Matt Markham, Deryk Anderson and Matt Warner for contributing their domain expertise and experience to the topic. The authors also acknowledge Ravi Kiran and Gnanananda Maditati for contributing their technical skills.

NOMENCLATURE

CMMS	Computerized Maintenance Management System
EAM	Enterprise Asset Management
BOW	Bag-of-words
RNN	Recurrent neural network
SVM	Support vector machine
NLP	Natural language processing
RF	Random Forest
RCM	Reliability-centered maintenance
FMEA	Failure mode and effects analysis
KPI	Key Performance Indicator
TTF	Time-to-failure
NHPP	Non-homogeneous Poisson Process
CBM	Condition-based Monitoring
RCA	Root cause analysis

REFERENCES

Abernethy, R. B. (2004). The New Weibull Handbook Fifth Edition, Reliability and Statistical Analysis for Predicting Life, Safety, Supportability, Risk, Cost and Warranty Claims. Spiral-bound Published and distributed by Robert B. Abernethy.

Arif-Uz-Zaman, K., Cholette, M. E., Li, F., Ma, L., & Karim, A. (2016). A data fusion approach of multiple maintenance data sources for real-world reliability modelling. *Proceedings of the 10th World Congress on Engineering Asset Management* . Sprint, Cham. 69-77.

Arif-Uz-Zaman, K., Cholette, M. E., Ma, L., & Karim, A. (2017). Extracting failure time data from industrial maintenance records using text mining. *Advanced Engineering Informatics* 33, 383-396.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine* 9 (2), 48-57.

Chollet, F. (2015). Keras. https://keras.io.

Distefano, R. &. Thomas, S. (2011) Asset Data Integrity is Serious Business. New York: *Industrial Press Inc*.

Dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics.* 69-78.

Fortin, J. B. (2018). Guidebook for Advanced Computerized Maintenance Management System Integration at Airports. No. Project 09-14, TRB's Airport Cooperative Research Program (ACRP).

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57, 345-420.

Gulati, R. (2013). Maintenance and Reliability Best Practices Second Edition. New York, NY: *Industrial Press, Inc*.

He, B. (2016). A Machine Learning Approach for Data Unification and Its Application in Asset Performance Management. Thesis, Virginia Polytechnique Institute and State University, Blacksburg, VA.

Hodkiewicz, M., & Ho, M. T. W. (2016). Cleaning historical maintenance work order data for reliability analysis. *Journal of Quality in Maintenance Engineering* 22 (2): 146-163.

Hodkiewicz, M., Batsioudis, Z., Radomiljac, T., and Ho, Mark T.W. (2017). Why autonomous assets are good for reliability - the impact of 'operator-related component' failures on heavy mobile equipment reliability. *Annual Conference of the Prognostics and Health Management Society.* 8.

ISO. 2004. *Petroleum and natural gas industries - Collection and exchange of reliability and maintenance data for equipment.* International Standards Organization (ISO), Genieve, Switzerland: International Standards Organization.

JASC. 2002. "Joint Aircraft System/Component Code Table and Definitions." Flight Standards Service Regulatory Support Division, Federal Aviation Administration, Oklahoma City, OK.

Konigstein, H. M. (2007). The RDS-PP - Transition from KKS to an international Standard. *GB PowerTech, 87*(8), 64-72.

Lukens, S., Naik, M., Doan, D., Hu, X. & Abado, S. (2017). The role of transactional data in prognostics and health management work processes. *Proceedings of the Annual Conference of the Prognostics and Health Management Society*. 517-528.

Lukens, S. & Markham, M. (2018). Data-driven application of PHM to asset strategies. *Proceedings of the Annual Conference of the Prognostics and Health Management Society.* Philadelphia, PA.

Lukens, S., Naik, M., Markham, M. & Laplante, M. (2019). Data-driven approach to estimate maintenance life cycle cost of assets. *Proceedings of the 10th Model Based Enterprise Summit*.

Meeker, W. Q. (1998). Statistical models for reliability data. New York: *John Wiley & Sons, Inc*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. &, Duchesnay, E. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825-2830.

Sexton, T. & Brundage, M. (2019). Standards needs for maintenance work order analysis in manufacturing. *Proceedings of the 2019 Model-Based Enterprise (MBE) Summit.* Gaithersburg, MD.

Sexton, T., Brundage, M. P., Hoffman, M., & Morris, K. C. (2017). Hybrid datafication of maintenance logs from AI-assisted human tags. Big Data (Big Data), IEEE International Conference.

Sexton, T., Hodkiewicz, M., Brundage, M. P., & Smoker, T. (2018). Benchmarking for Keyword Extraction Methodologies in Maintenance Work Orders. *Proceedings of the Annual Conference of the Prognostics and Health Management Society*.

Sexton, T., Moccozet S., Brundage, M. P., Madhusudanan, N., Hastings, E., & Bones, L. (2018) Nestor: Quantifying tacit knowledge for investigatory analysis, Github, https://github.com/usnistgov/nestor.

Sikorska, J., Hammond, L., & Kelly, P. (2008). Identifying failure modes retrospectively using RCM data. *ICOMS Asset Management Conference.* Melbourne, Australia.

BIOGRAPHIES

Kittipong Saetia is a Data Scientist at GE Digital in Product R&D group. His role involves researching and developing data analytics solutions for Asset Performance Management and Operation Performance Management applications. His research interests are machine learning, natural language processing, and the Industrial Internet of Things. He received his M.S.CEP and Ph.D. in chemical engineering in 2013 from the Massachusetts Institute of Technology. Prior to joining GE in 2017, he worked at Intel Corporation in Oregon as a process development engineer responsible for leading scientific research to improve manufacturing yield of next-generation processor technology.

Sarah Lukens lives in Roanoke, Virginia and is a Data Scientist at GE Digital. Her interests are focused on data-driven modeling for reliability applications by combining modern data science techniques with current industry performance data. This work involves analyzing asset maintenance data and creating statistical models that support asset performance management (APM) work processes using components from natural language processing, machine learning, and reliability engineering. Sarah completed her Ph.D. in mathematics in 2010 from Tulane University with focus on scientific computing and numerical analysis with applications in fluid-structure interaction problems in mathematical biology. Prior to joining Meridium (now GE Digital) in 2014, she conducted post-doctoral research at the University of Pittsburgh and the University of Notre Dame building data-driven computational models forecasting infectious disease spread and control. Sarah is a Certified Maintenance and Reliability Engineer (CMRP).

Henri (Erik) Pijcke is a Services Engineer with GE Digital who works with industrial maintenance organizations collecting and creating actionable insights from their CMMS data. Erik has been implementing CMMS and APM systems for the last 30 years and believes that many maintenance

organizations are trapped in a reactive "firefighting" mode. Due to their corporate equipment hierarchy and maintenance work processes that are mostly focused on efficiently executing reactive work, they are unable to shift from a reactive to a proactive maintenance organization. Carefully analyzing the CMMS data uncovers maintenance work processes and activities that prohibit effective reliability studies. Erik holds a Bachelor's Degree in Mechanical Engineering from the HTS in Vlissingen, the Netherlands.

Xiaohui (Mark) Hu is currently principal data scientist at GE Digital, located in Boston, MA. He is working in Product R&D group to support multiple software product lines within Predix Asset Performance Management and Operation Performance Management family. Prior to joining GE, Mark was assistant research professor at Purdue University at Indianapolis. Mark received his PhD in Electrical and Computer Engineering from Purdue University and his Bachler degree from Tsinghua university, China respectively. His main research interests are machine learning, prognostics and health management, swarm intelligence, computational intelligence, optimization, and data modeling. He is a senior member of IEEE and serves as reviewers committee members in multiple international journals and conferences.