

SEMANTIC TEXT ANALYTICS TECHNIQUE FOR CLASSIFICATION OF  
MANUFACTURING SUPPLIERS

by

Ramin Sabbagh, B.Sc.

A thesis submitted to the Graduate Council of  
Texas State University  
In partial fulfillment of the requirements for the degree of  
Master of Science  
With a Major in Technology Management  
May 2018

Committee Members:

Farhad Ameri, Chair

Jaymeen Shah

Bahram Asiabanpour

**COPYRIGHT**

by

Ramin Sabbagh

2018

## **FAIR USE AND AUTHOR'S PERMISSION STATEMENT**

### **Fair Use**

This work is protected by the Copyright Laws of the United States (Public Law 94-553, section 107). Consistent with fair use as defined in the Copyright Laws, brief quotations from this material are allowed with proper acknowledgement. Use of this material for financial gain without the author's express written permission is not allowed.

### **Duplication Permission**

As the copyright holder of this work I, Ramin Sabbagh, refuse permission to copy in excess of the "Fair Use" exemption without my written permission.

## **DEDICATION**

I would like to dedicate my thesis to my wife and my family. My wife Paria has always been there for supporting me and encouraging me through all the challenges of graduate life. I am truly thankful for having her in my life. I want to extend my feeling of gratitude towards my Dad for being my inspiration and my Mom for her endless love, and care. I also want to dedicate my thesis to my Brothers Mohammad Amin and Mohammad Mahdi, who never left my side and always cheered me up.

## **ACKNOWLEDGEMENTS**

I have no words to express my gratitude towards Dr. Ameri for being such a wonderful supervisor. I am thankful to him for giving me the opportunity to work in his lab. I consider myself blessed to have him as my advisor. He always believed in me and encouraged me to perform my best. He will always inspire me by being an outstanding teacher, a great researcher and above all a great human being.

I would like to thank my thesis committee members Dr. Shah for enlightening me the glance of research and Dr. Asiabanpour for his valuable insights and guidance.

I would also like to thank Dr. Aguayo and Dr. Torres for their support, stimulating discussions and all the suggestions they provided to accomplish my research. I thank my fellow lab mates in Texas State for being with me all the time.

Finally, I want to recognize that this research would not have been possible without the assistance of Texas State University, Department of Engineering Technology at TXST, and the National Institute of Standards and Technology (NIST) and express my gratitude to them. Funding for this research is provided by the National Institute of Standards and Technology (NIST) under collaborative agreement 70NANB14H255.

## TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS .....	v
LIST OF TABLES .....	xi
LIST OF FIGURES .....	xiii
LIST OF ABBREVIATIONS .....	xvi
ABSTRACT.....	xvii
CHAPTER	
I. RESEARCH PERSPECTIVE.....	1
1. Background and Motivation .....	1
2. Problem Statement .....	3
3. Research Question .....	4
4. Assumption, Limitation, and Delimitation .....	5
4.1. Assumption: .....	5
4.2. Limitation:.....	5
4.3. Delimitation: .....	5
5. Related Works .....	6

6. Research Methodology .....	8
7. Research Plan .....	11
II. MANUFACTURING SUPPLIER CLASSIFICATION .....	12
1. Introduction .....	12
1.1. Approaches to Manufacturing Capability Modeling .....	13
1.2. Text Analytics for Capability Data Mining .....	14
2. Related Work in Capability Modeling .....	16
3. Industrial Applications of Text Mining.....	20
4. Manufacturing Capability Thesaurus (MCT).....	24
4.1. SKOS .....	25
4.2. Thesaurus Development and Extension Process.....	27
4.3 Capability Model in MCT.....	30
5. Manufacturer Classification Framework.....	33
5.1. Bag-of-Concepts (BOC) Instead of Bag-of-Words (BOW) Method .....	34
5.2. Naïve Bayes Method.....	34
5.3. Concept Model Learning .....	36
5.4. Test Document Classification .....	39

6. Implementation and Experimental Validation .....	40
6.1. Class Definition.....	40
6.2. Generating Concept Models for Heavy and Complex Machining Classes .....	42
6.3. Test Data Preparation .....	45
6.4 Test Data Conceptualization .....	47
6.5. Performance Evaluation .....	49
6.6. Results .....	50
7. Conclusion.....	54
III. MANUFACTURING SUPPLIER CLUSTERING .....	55
1. Introduction .....	55
2. Related Work in Unsupervised Learning .....	56
3. Overview of the Framework.....	60
3.1. Extracting the Raw Text from Manufacturing Supplier's Website .....	61
3.2. Preprocessing .....	61
3.3. Extracting N-grams after Preprocessing .....	61
3.4. TF-IDF Normalization on Term Frequency.....	62



3.5. Identifying the Useful N-grams .....	63
3.6. K-means Clustering.....	66
3.7. Topic Modeling.....	68
4. Experiment .....	69
4.1. Dataset Preparation .....	70
4.2. Data Preprocessing.....	71
4.3. Extracting N-grams from Supplier.....	71
4.4. Normalization Based On TF-IDF .....	72
4.5. Detecting Meaningful N-grams from Heavy Machining Suppliers .....	73
4.6. K-Means Clustering Result on Suppliers of Heavy Machining and Complex Machining Classes .....	75
4.7. Clustering Result of Manufacturing Suppliers from ThomasNet with No Categorical Preference .....	76
4.8. Topic Modeling Results .....	80
5. Conclusion.....	82
IV. CONCLUSION AND FUTURE WORK .....	84
1. Answers to Research Questions .....	84

2. Contribution.....	91
2.1. Methodological Contributions .....	91
2.2. Information Models and Tools.....	92
3. Future Work .....	93
APPENDIX.....	95
REFERENCES .....	119

## LIST OF TABLES

Table	Page
1. Manufacturing Capability Criteria Proposed by Lekurwale and Braker .....	18
2: The Concepts under the Manufacturing Capability Concept Schema.....	31
3: Concept Extraction via SPARQL Queries.....	37
4: Entry Concepts for Target Classes.....	41
5: Some of the Members of Complex Machining and Heavy Machining Concept Models.....	44
6. Concept Vectors for 10 Suppliers Belonging to Heavy Machining and Complex Machining Classes .....	48
7: Data Required for Calculation of Precision, Recall, and F-measure for Heavy Machining .....	50
8. Detail Result of the Two Scenarios and Four Classification Techniques.....	52
9. T-Test to Compare the Result of Scenario One and Two for Different Classification Techniques .....	53
10. T-Test to Compare the Result of Bag of Words and Bag of Concepts Approach for Naïve Bayes Classification Techniques.....	53
11. Top 10 Candidate N-grams resulted from the Total Frequency Analysis .....	74
12. Top 25 Most Important N-Grams for Heavy Machining Suppliers Resulted by Latent Semantic Analysis Method .....	75

13. K-means Results on Heavy Machining and Complex Machining Classes .....	76
14. Top 20 Most Important N-Grams of Unseen Manufacturing Suppliers Dataset	
Resulted by Latent Semantic Analysis (LSA) Method .....	77
15. Topic Modeling Results for 150 Unseen Manufacturing Suppliers .....	81
16. Topic Modeling Result on the Cluster Related to Casting Suppliers .....	82

## LIST OF FIGURES

Figure	Page
1. The SKOS Concept Diagram for Swiss Machining Process .....	27
2. Document Frequency of Some of the Concepts (Categorized Based on the Schema) in an Intermediate Stage of Thesaurus Development (Before Deleting Less Frequent Concepts) .....	29
3: Total Number of Concepts under Each Concept Scheme in the Manufacturing Capability Thesaurus .....	32
4: Total Number of Concepts under Manufacturing Capability .....	33
5: Proposed Manufacturer Classification Framework .....	35
6: Concept Weighting Schema.....	38
7. Concept Model Builder Function.....	42
8.The User Interface for Extracting Capability Text .....	45
9. Sample Capability Narrative Tagged by MC Thesaurus Concepts .....	46
10. Extracted Concepts and Frequencies of the Sample Capability Narrative .....	47
11. The Precision of the Text Classification .....	51
12. Bag of Concepts VS Bag of Words .....	51
13. Major Steps to Find Meaningful N-grams .....	65
14. Main Steps of the Clustering Algorithm.....	67
15. Plain Text from an Example Supplier in the Dataset.....	70

16. Example Supplier after Preprocessing .....	71
17. Bigram Terms Detected from Example Supplier.....	72
18. Trigram Terms Detected from Example Supplier .....	72
19. Partial View of the Preprocessed Data before Applying TF-IDF for the Example Supplier .....	73
20. A partial View of the Dataset after Applying TF-IDF Function for the Example Supplier .....	73
21. Relationships between Number of Clusters and Within Groups Sum of Squares .....	78
22. The Result Map of the K-means Clustering Technique for 150 Suppliers Based on <i>Casting</i> and <i>Machining</i> Normalized Frequencies.....	79
23. The Result Map of the K-means Clustering Technique for 150 Suppliers Based on <i>Assembly</i> and <i>Forming</i> Normalized Frequencies .....	80
24. Home Page of SKOSTool Website.....	95
25. Screenshot from Selection of the Entry Concepts .....	97
26. Partial View of the Important Concepts of Complex Machining Class.....	97
27. Ability to Delete One or More Concepts from Concept Model.....	98
28. Modifying the Concepts inside the Concept Model .....	99
29. Adjusting the Weightings inside the Concept Model .....	100
30. Input Text.....	101
31. URL Preview .....	102
32. Analysis Result .....	103

33. Vector Model from an Example supplier.....	104
34. R Codes for Bag of Word Approach-Naive Bayes Technique .....	106
35. R Codes for Bag of Word Approach-Naive Bayes Technique - Continued .....	107
36. R Codes for Bag of Concept Approach - Naïve Bayes Techniques .....	108
37. R Codes for Bag of Concept Approach - KNN Techniques .....	108
38. R Codes for Bag of Concept Approach - Random Forest Techniques .....	109
39. R Codes for Bag of Concept Approach - Random Forest Techniques .....	109
40. Unsupervised Learning – First Part .....	111
41. Unsupervised Learning – Second Part.....	112
42. Unsupervised Learning – Third Part.....	113
43. Unsupervised Learning – Forth Part .....	114
44. Unsupervised Learning – Fifth Part.....	115
45. Unsupervised Learning – Sixth Part .....	116
46. K-Means Clustering .....	117
47. Topic Modeling Technique.....	118

## LIST OF ABBREVIATIONS

MCT .....	Manufacturing Capability Thesaurus
SKOS .....	Simple Knowledge Organization System
P .....	Precision
R.....	Recall
F .....	F-measure
HM .....	Heavy Machining
CM .....	Complex Machining
SVM.....	Support Vector Machine
KNN .....	K-Nearest Neighbor
BOW .....	Bag of Words
BOC .....	Bag of Concept
EE.....	Entity Extractor
LSA.....	Latent Semantic Analysis



## **ABSTRACT**

Most of the information available in the manufacturing industry is in unstructured, natural language format. The unstructured data could contain important and useful information that can inform decision makers across different phases of product lifecycle. However, due to its unstructured nature, it is often difficult to effectively use the information embedded in the data represented in plain text. Manufacturing Capability data is one type of data often represented in unstructured format on the websites of manufacturing firms. If manufacturing capability data is parsed, organized, and analyzed properly, it can be used for supplier evaluation and selection during supply chain formation process. In order to come up with an efficient method of capability analysis, it is important to identify the main characteristics of the capability. Different aspects of manufacturing capability include manufacturing processes, industry coverage, engineering, organizational, and quality capabilities. There are several methods that can be used for extracting information from text. Data mining is one of the most powerful methods which is currently used for different knowledge extraction purposes.

This research presents a method for manufacturing capability analysis and modeling through implementation of different supervised and unsupervised text mining methods using unstructured text in suppliers' website as the input. For supervised text mining, Naïve Bayes, KNN, SVM, and Random Forest methods are used as the analytical classification techniques. The objective is to classify suppliers into pre-labeled classes

based on the textual description of their capabilities. In unsupervised text mining method, two popular methods, namely, Clustering and Topic Modeling methods are used to split the diverse suppliers into several groups and then, find the appropriate characterizations associated with each group. The proposed methods are evaluated experimentally using real capability data collected from the webpages of manufacturers in contract machining industry. In order to evaluate the accuracy of the results, precision, recall, and F-measure are used as the metrics.

# CHAPTER I

## RESEARCH PERSPECTIVES

### 1. Background and Motivation

Manufacturing sector has one of the fastest rate of data generation among all sectors of the economy. Most of the information generated by various sources in manufacturing is in unstructured form. Unstructured data, such as natural language text that tends to be human-generated, is a type of data that does not conform to a specific, pre-defined data model and cannot be organized in a pre-defined manner. There are diverse examples of unstructured data in manufacturing industry. Most manufacturing documents, such as quality reports, maintenance logs, requirement specifications, or company blogs and whitepapers are instances of unstructured information. This thesis is mainly focused on manufacturing capability data. There is a high volume of unstructured, but useful, data in suppliers' websites which could contain very important information about different aspects of manufacturing capability such as process, material, industry, geometry, quality, and engineering capabilities.

One of the common challenges in dealing with unstructured data is data cleansing and preparation for the sake of extracting information from data. It involves reducing the dimensionality of data and cleaning the noise features. It is also necessary to have some provisions for locating, extracting, organizing, and storing the unstructured data. Second challenge is data volume. The amount of unstructured data increases continuously and it is essential to come up with most updated data and useful information associated with it.

Another challenge is data relevance. It is difficult to find the relevant resources for target purposes.

This research is motivated by the need for supporting supply chain decisions by supplier capability information. In manufacturing supply chain management, it is essential to gain adequate insight into different aspects of suppliers' capabilities in order to make more informed decisions when selecting manufacturing partners. Manufacturing suppliers could benefit from this capability information and identify pros and cons of the competitor manufacturing suppliers. On the other hand, customers could utilize the information about suppliers to compare multiple suppliers based on their important characteristics. In order to conduct a reliable evaluation and comparison, it is critical to have legitimate criteria associated with manufacturing and supply chain management concepts.

It is required to gather information about all of these capabilities as a means to recognize a specific supplier's type and properties. In order to benefit from unstructured information and convert multiple manufacturing capability to valuable information and knowledge, it is required to have an appropriate methods, techniques, and tools. Data mining is one of the most practical and effective tools which is currently used for different knowledge extraction purposes. Since we are dealing with raw text of suppliers' website, one of the special types of data mining, namely, text mining, is used for capability modeling and ranking of suppliers.

There are five major types of data mining methods, namely, anomaly detection, association rule learning, clustering analysis, classification analysis, regression analysis, and summarization. In this research, two popular methods, namely, classification analysis

and clustering analysis is used. And based on these two methods, several techniques such as K-Nearest Neighbor, Naïve Bayes, Support Vector Machine, Random Forest, Topic Modeling, methods is used for capability analysis and information extraction. These techniques contain supervised and unsupervised learning as well as phrase mining methods.

Data classification (James, 1985) is the process of organizing data into categories for its most effective and efficient use. A well-planned data classification system makes essential data easy to find and retrieve. In this method, some predefined classes should be defined based on required information. After definition of the classes, it is essential to find critical criteria associated with each class. After assigning the criteria and specific properties to the class, each supplier near those criteria, is classified under that specific class.

In data clustering, the classes are not pre-defined and the goal is to split the diverse documents and suppliers into several groups which have the common characteristics and then, find the appropriate characterizations and specifications associated with these groups (Gan, Ma, & Wu, 2007; Jain, Murty, & Flynn, 1999; T. Zhang, Ramakrishnan, & Livny, 1996).

## **2. Problem Statement**

There is high volume of useful information hidden in suppliers' websites. In manufacturing area, suppliers provide valuable information about their capabilities such as the product they manufacture the processes they offer, and qualities they can achieve. Capability data published on suppliers' websites is often in unstructured format. Several

approaches such as machine learning and data mining can be applied to organize the unstructured data and make it more usable. If manufacturing capabilities can be analyzed and evaluated based on the textual information provided on suppliers' websites, more informed decisions can be made when forming manufacturing supply chains. To achieve this goal, there is a need for development of an automated text mining tool supported by analytical techniques. The *objective of this research* is to create a capability analysis framework for manufacturing supplier's classification through implementation of different supervised and unsupervised text mining techniques.

### **3. Research Question**

This research work is intended to answer the following questions:

- What is the most suitable text classification technique for capability-based supplier classification problem?
- What are the steps needed for data preparation and cleaning?
- How to manually create the corpus and manipulate it as training data in data mining techniques?
- How text classification techniques can help organize suppliers based on their capabilities?
- What types of hidden knowledge patterns exists in manufacturing suppliers' websites?
- What are the important terms which suppliers, in contract manufacturing industry, use in order to describe their capabilities?

#### **4. Assumption, Limitation, and Delimitation**

The assumptions, limitations, and delimitations of the research is provided in this section.

##### **4.1. Assumption:**

- All the information provided in the supplier's website reflects the true capabilities of that supplier.
- Each supplier could belong to several classes. For example, a specific supplier could belong to heavy machining class, as well as medical industry class.
- The thesaurus which is used in the research is semantically valid.
- Thomasnet.org or similar websites are used as the data source. They are assumed to have correct supplier categorization in their website.

##### **4.2. Limitation:**

- Capability information is only obtained from the public websites of suppliers and the internal capability information is not available to be included in the analysis.
- Number of suppliers representing different classes are limited and varies between two specific classes. This, makes it difficult to collect appropriate and adequate training and test data.

##### **4.3. Delimitation:**

- This study only covers manufacturing suppliers.
- The suppliers that is studied in this thesis belong to contract machining industry. Other processes, such as assembly, forming, etc. are excluded from this study.

## 5. Related Works

As a consequence of ascending growth of data, especially in the internet, users, organizations, and companies tend to utilize knowledge extraction tools and methods such as machine learning, data mining, etc. Accordingly, researchers tried to employ multiple applications of knowledge extraction methods to manufacturing industry and supply chain management domain.

Wang et al. (F. Wang, Wang, Li, & Wen, 2014) came up with a method which uses bag of concepts instead of traditional bag of terms technique aiming to resolve the surface mismatching and polysemy problems. First, the concept model is created for each target category through enabling a large knowledgebase such as thesaurus. Then, the concept-based similarity mechanism can classify phrases and short text to the most similar category. The bag of concept method facilitates text ranking after the classification.

In order to classify the enterprise websites, Dong and Liu (Dong & Liu, 2006) presented the website topic feature modeling method using support vector machine techniques. They utilized a multi-feature topic vector generated by the website's textual content as well as content structure to determine the genre of website. The proposed method was verified by conducting an experiment on manufacturing enterprise website search.

Lee and Hong (Lee & Hong, 2016) proposed a data-driven method to extract industrial service status from companies' annual reports using text mining algorithm. The method automatically recognized the word-usage patterns and evaluate the service status.



Employing self-organizing map, the major service clusters alongside niche areas in the market was identified which are two major parameters to service development planning.

Wang (K. Wang, 2007) discussed the nature and significance of data mining techniques in manufacturing. He explained the implementation of data mining on product design and manufacturing and how the productivity and efficiency of manufacturing suppliers and companies can be influenced by product design, manufacturing process, decision making, and management.

Harding et al. (Harding, Shahbaz, Kusiak, & others, 2006) reviews applications of data mining in manufacturing engineering in multiple subjects such as decision support, fault detection, product quality improvement, etc. He also discussed information integration aspects, customer relationship management, and standardization.

Shotorbani et al. (Shotorbani, Ameri, Kulvatunyou, & Ivezic, 2016) proposed a method using clustering and topic modeling to improve searching and organizing textual documents and extract valuable patterns from manufacturing websites. The method demonstrated that topic modeling along with document clustering, boost annotation and classification of manufacturing supplier's webpages. It assisted users to extract valuable patterns from supplier's websites.

Jung et al. (Jung, Kulvatunyou, Choi, & Brundage, 2016) proposed a method for assessing factories' readiness for implementation of technologies which can be served to create smart manufacturing systems. The method evaluates the companies and provide users with the status of current target company's readiness level in comparison to the reference model. Knowing the current state, companies can intend to improve their

readiness level which is verified that has a positive correlation with companies' operational functions.

Salami et al. (Salami, TaghaviFard, & Majidifar, 2015) arranged technological capability assessment indicators in order to make formulation of future policies and strategic planning more productive. In similar research, Cheraghi et al. (Cheraghi, Dadashzadeh, & Subramanian, 2004) presented the critical success factors in order for select the supplier based on its rank. They realized that supplier selection criteria have changed during passing of time and will change in the future.

Most of the knowledge extraction methods in the literature review are traditional text mining methods and using plain text as an input to their experiments. Bag of concept methods and thesaurus based capability analysis and text mining is one of the novel methods which has been used in manufacturing domain.

## **6. Research Methodology**

For supervised classification, the Bag of Concepts (BoC) method is used. In traditional text mining technique (i.e. bag of terms) the term dictionary is built automatically with different approaches such as machine learning, using the most frequent terms of the documents in the corpus. One of the disadvantages of this method is breaking multi-term phrases into single words resulting in semantic loss. As an alternative method, the bag of concepts method is presented in this research which preserves the meanings of concepts and phrases. The concepts are directly used for tokenizing textual documents.

In the proposed method, it is required to manually create a dictionary of manufacturing concepts. Accordingly, the manufacturing capability thesaurus (MCT) is created and the concepts associated with manufacturing capability are organized and linked using a set of semantic relationships. Six concept schemes, namely, manufacturing capability, organizational capability, engineering capability, quality capability, industry capability, and general capability are defined. Multiple manufacturing capability concepts are added under those concept schemes in order to better organize the thesaurus and boost the power of analyzing the manufacturing texts.

Once the thesaurus contains adequate concepts under different concept schemes, the target classes are defined and the relevant concepts associated with target classes are identified and a weighting is assigned to each concept which demonstrates the importance of specific concept in specific target class through. The process of identifying important concept and assigning the weightings are done in Concept Model Builder section of the SKOSTool website that is described in Chapter 2. The SKOSTool is web-based tool programmed by JAVA in INFONEER lab<sup>1</sup> in order to facilitate the process of text analytics. Some of the important capabilities associated with the SKOSTool are analyzing the corpus of text data, creating concept models for target classes, and detecting the frequencies of concepts in the websites.

Each supplier website's text serves as raw data which is used to find the frequencies of the candidate concepts that are already identified in concept model in each supplier. For this purpose, the Entity Extractor (EE) tool is created in SKOSTool in order

---

<sup>1</sup> <http://infoneer.wp.txstate.edu/>

to automatically extract candidate concepts' frequencies from multiple suppliers and provides user with website's raw text as well as CSV format of concepts and associated frequencies.

Once the concept model as well as supplier data are available, the experiments could be run in two different scenarios. Finally, four frequently used techniques, namely, Naïve Bayes, K-Nearest Neighbor or KNN, Support Vector Machine or SVM, and Random Forest Methods are applied into both dataset from both scenarios. Finally, in order to evaluate the accuracy of the methods, three popular measures, namely, precision, recall, and F-measure are calculated at the end of experiments.

In unsupervised text mining, the classes are not defined beforehand and the goal is to split the diverse documents and suppliers into several groups and then, find the appropriate characterizations and specifications associated with each group. The data in unsupervised learning experiment is collected through suppliers' website. Several preprocessing functions are applied in order to make a clean dataset. The Latent Semantic Analysis (LSA) is adopted in order to detect the most significant terms and documents of the dataset. Two popular methods, namely, Clustering Method and Topic Modeling Method are also used in order to explore the dataset and detect the characteristics of the dataset. Manufacturing suppliers in both supervised and unsupervised learning methods are collected from ThomaNet.com and MFG.com.

## **7. Research Plan**

The work included for this research is broken down into 7 tasks as listed below:

- Task 1: Literature Review on Text Mining
- Task 2: Modeling and Implementation
  - Task 2.1: Supervised
  - Task 2.2: Unsupervised
- Task 3: Thesaurus Development
- Task 4: Data Collection
- Task 5: Experiment and Analysis

## CHAPTER II

### MANUFACTURING SUPPLIER CLASSIFICATION

#### 1. Introduction

Capability analysis is a necessary step in the early stages of supply chain formation. Most existing approaches to manufacturing capability evaluation and analysis use structured and formal capability models as input. However, manufacturing suppliers often publish their capability data in an unstructured format. The unstructured capability data usually portrays a more realistic view of the services a supplier can offer. If parsed and analyzed properly, unstructured capability data can be used effectively for initial screening and characterization of manufacturing suppliers specially when dealing with a large pool of prospective suppliers.

This work proposes a novel framework for capability-based supplier classification that relies on the unstructured capability narratives available on the suppliers' websites. Naïve Bayes is used as the text classification technique. One of the innovative aspects of this work is incorporating a thesaurus-guided method for feature selection and tokenization of capability data. The thesaurus contains the informal vocabulary used in the contract machining industry for advertising manufacturing capabilities. An Entity Extractor Tool (EE) Tool is developed for the generation of the concept vector model associated with each capability narrative. The proposed supplier classification framework is validated experimentally through forming two capability classes, namely, heavy component machining and difficult and complex machining, based on real capability data.

The manufacturing industry is undergoing profound changes brought about by the emergence of service-oriented, cloud-based, and digital manufacturing paradigms. The democratization of manufacturing is among the most visible trends that have reshaped the manufacturing landscape within the past few years. With a lowered barrier to entry, a larger number of small-to-medium sized enterprises (SMEs) are capable of offering diverse manufacturing services through building virtual supply networks and exploiting the resources provided by distributed partners. Consumers of manufacturing services can benefit from a larger and more diverse supply pool since they are provided with a wider range of options when searching for qualified suppliers. Nevertheless, the sheer size of the supply pool presents multiple challenges to efficiently evaluating and selecting manufacturing suppliers. Traditional approaches to supplier evaluation and selection often entail direct interaction with the supplier and possibly visiting the supplier's facility to obtain better insight into the technological and organizational capabilities of the supplier. However, as the interaction between suppliers and customers becomes increasingly virtual and the lifespan of supply chains becomes shorter, more agile and data-driven approaches to capability evaluation are called for. This research is motivated by the need for improving the agility and intelligence of supplier discovery and evaluation solutions and also enhancing the visibility of SMEs in the cyber-space.

### **1.1. Approaches to Manufacturing Capability Modeling**

Manufacturing capability modeling and representation has been addressed by multiple researchers using formal, standard, and structured modeling approaches (Ameri & Sabbagh, 2016; Lin et al., 2016). The approaches that are based on structured capability models are effective in centralized scenarios, such as e-sourcing portals, where

SMEs are registered with the portal and their capability models are maintained in a single repository that is governed by a predefined reference model. A centralized scenario creates a controlled environment which is more amenable to accurate and in-depth capability modeling and quantification. The caveat of the centralized approach is that the search space is limited to only those suppliers who have chosen to register with the portal. The smaller supply pool could lead to the formation of sub-optimal supply chains regardless of the level of sophistication and rigor incorporated in the underlying search and evaluation algorithms. In decentralized scenarios, on the other hand, the search space is extended to the entire web. The SMEs use their websites to advertise their services and capabilities more thoroughly without being restrained by predefined templates. However, since SMEs are not forced to comply with predefined controlled vocabularies, standard reference models, and rigid data structures, the manufacturing capability information is highly unstructured, heterogeneous, and ill-defined. For this reason, the keyword search method using generic search engines cannot efficiently narrow down the search space to a set of highly relevant suppliers. More sophisticated text analytic techniques are required for processing, organizing, and analyzing the unstructured capability data and forming capability classes with known properties.

## **1.2. Text Analytics for Capability Data Mining**

There are different text analytics techniques that can be applied to capability data, including classification, clustering, and topic modeling. This research focuses on the text classification problem. In particular, the objective of this research is to introduce a methodology for *automated classification of manufacturing suppliers* based on their capability data provided in natural language. One advantage of the capability narrative



directly provided by the manufacturer is that it depicts a more realistic picture of the manufacturer's capabilities since it is not constrained by predefined templates and vocabularies.

In text classification problems, one of the key steps is selecting the features (i.e., terms and/or concepts) that can uniquely characterize a class of interest. *Feature selection* (a.k.a feature engineering) can be done either manually or automatically through machine learning. Manual feature selection entails manual selection of the features for each class by domain experts. Manual feature engineering often requires considerable effort, and the selected features might be subjective and problem-specific. Automated feature engineering requires less time and effort compared to the manual method but its main drawback is that it often ignores the semantic dependencies between the features (Saeys, Inza, & Larrañaga, 2007).

In this work, a semi-automated technique, guided by a formal thesaurus, is adopted for feature engineering. The formal thesaurus developed in this work uses SKOS (Simple Knowledge Organization system) (Miles & Bechhofer, 2009) syntax and semantics and contains the terms used in the contract manufacturing industry for describing their capabilities. The particular focus of the thesaurus is on the CNC machining industry.

The thesaurus was built in a bottom-up fashion through tagging the relevant terms in a training corpus and connecting them together through semantic and lexical relationships. Different subsets of the terms in the manufacturing capability thesaurus can be used as the distinguishing features for each capability class. Based on the proposed classification methodology, the unstructured capability narrative for each manufacturer is

converted into a vector of terms that is then fed into a text classifier. The text classifier used in the proposed framework uses the Naïve Bayes algorithm (Murphy, 2006). In the presence of a thesaurus, one can dynamically define various capability classes simply by selecting the relevant terms, or features, available in the thesaurus for the class of interest. The feature selection process is guided by the underlying semantic model of the thesaurus. This eliminates the time-consuming and costly steps required for creating gold standard training corpus for each class of interest that is often used in supervised classification techniques based on machine learning. The formal nature of the proposed thesaurus enhances the semantic relevance of the results.

The proposed classification methodology is validated experimentally through forming two capability classes, namely, heavy component machining and difficult and complex machining. The standard Information Retrieval (IR) metrics, such as precision, recall, and F-measure are used to evaluate the performance of the classifier.

The remainder of this paper is organized as follows. In Section 2, the related work in manufacturing capability modeling is discussed. Section 3 introduces the Manufacturing Capability Thesaurus (MCT). The proposed classification framework is discussed in Section 4 and the results of an experimental validation of the framework are provided in Section 5. The paper ends with the conclusions.

## **2. Related Work in Capability Modeling**

Manufacturing capability can be defined as the “firms’ internal and external organizational skills, resources, and functional competencies to meet the requirement of the changing economic environment” (Teece, 1990). Manufacturing capability is a multi-

faceted entity represented through criteria such as quality, processing capability, production capacity, flexibility, product innovation capacity, and performance history.

Different researchers have introduced various indicators and metrics for manufacturing capability and used different quantitative techniques for capability measurement and assessment. The objective of this literature study is to identify the primary indicators of manufacturing capability and to inform the processes of forming a taxonomy of formal and informal capability-related terms to be used in the thesaurus.

Hayes et al. (Hayes & Wheelwright, 1984) considered capacity, facilities, process technologies, vertical integration and vendors, human resources, quality, production planning, new product development, performance measurement and reward, organization and system as ten major capability assessment factors. Skinner (Skinner, 1969) used plant and equipment, product engineering, labor and staffing, production planning and control, and organization and management as the primary criteria for evaluating manufacturing capability. Miltenburg (Miltenburg, 2005) considered human resources, organization structure, sourcing, production planning and control, process technologies, facilities as six major capability indicators. Nigel and Michael (Nigel & Michael, 2008) recommended capacity, process technologies, supply network, organization and development as four major manufacturing capability criteria.

Liu et al. (Liu, Jiang, & Cao, 2014) considered processing capability and production capability as the most significant criteria of manufacturing capability. They proposed four major factors for process capability, namely, processing material, dimension, type & feature, and accuracy.

Luo et al. (Luo et al., 2013) used a multi-dimensional method to evaluate manufacturing capability based on fuzzy information and dynamic behavior description. Their conceptual model of manufacturing capability consists of four factors: Resources (such as software, hardware, and human), Task (such as quality, time, and cost), Process (such as model, method and flow) and knowledge.

**Table 1. Manufacturing Capability Criteria Proposed by Lekurwale and Braker**

Capability Factors (Lekurwale)		Capability Factor (Braker)
Skill level	Setup to run time	Operation cost
Nature of job	Scheduling uncertainty	Processing time
Performance appraisal	Production information required	Setup time per part
Training need	Length of planning for finish goods	Number of defects
Employee participation	Batching of backlog for planning	Time for producing defect
Wage rate	Type of layout	Labor time
Work content	Degree of automation	Labor cost
Decision making	Type of tooling	Part similarity
Organization structure	Use of AMT for process design	Workstation usage
Importance of line staff	Degree of coupling	Nominal part lead time
Quality responsibility	Degree of vertical integration	Material handling distances
Planning strategy	Material requirement prediction	Material handling time
New material inventory	Number of suppliers	Material handling cost
WIP inventory	Control over suppliers	Cost of defect production
Finish good inventory	Relationship with suppliers	Workstation down time
Planning input	Size of facility	Type of facility

Lekurwale et al. (Lekurwale, Akarte, & Raut, 2015) introduced a capability quantification method based on a multi-criteria approach. To quantify the manufacturing capability, he came up with 33 criteria based on studying the available literature review. He classified these 33 detailed criteria into six major categories, namely, human resources, organization structure and control, process technology, production planning and control, facilities. Using analytical hierarchy process (AHP), Lekurwale created a manufacturing capability model to quantify manufacturing capability. The capability level is then mapped into four abstract levels, namely, infant, industry average, adult, and world class as proposed by the Miltenburg (Miltenburg, 2005).

Baker and Maropoulos (Baker & Maropoulos, 1998) used a methodology for capability analysis for different abstraction levels, namely, cell, product, part, workstation and operation. Manufacturing capability criteria which are proposed by Lekurwale and Braker and Maropolous are listed in table 1. Some of the capability indicators listed in this table, such as downtime, defect rate, and utilization rate, are directly linked to the operational data that is often considered to be proprietary data and not available publicly. The proposed capability-based supplier classification method uses the publicly available information shared on the company's website. This study also revealed that most of the manufacturing capability criteria used by different researchers can be classified under a few board categories such as processing, quality, capacity, automation, geometric, and material capabilities. They can be treated as the main buckets of capability terms when building the thesaurus.

### **3. Industrial Applications of Text Mining**

The role of information management has become important in recent years. The ability to quickly find the specific and detail information from large amount of raw data is an approach for companies to pass and beat the competitors. Most of the information available in the manufacturing industry is in unstructured, natural language format. The unstructured data could contain important and useful information that can inform decision makers across different phases of product lifecycle. However, due to its unstructured nature, it is often difficult to effectively use the information embedded in the data represented in plain text. In recent year, text mining has become an important area of research in several fields, especially in industrial related domains.

Ur-Rahman and Harding (Ur-Rahman & Harding, 2012) proposed a hybrid methodology for classifying Post Project Review (PPR) documents into good and bad classes. The good documents are the ones that contain useful information in the context of their work. The first step of the proposed methodology involves using clustering techniques to partition the dataset into similar clusters. The second step results in generation of key phrases related to each cluster. They used A-priori Rule Association Mining technique to generate Multiple Key Term Phrasal Knowledge Sequences (MKTPKS) for each cluster. The third step deals with classifying the textual data into predefined classes. Different classification algorithms, such as KNN, Support Vector Machine, and Nave Bayes, were applied to test the performance of the proposed methodology. Through a case study based on industrial documents, they demonstrated that using a multiple-term approach provides a better accuracy compared to single-term

approaches. It is not clear how the identified multi-terms are validated before being used for classification purpose.

Huang and Murphey (Huang & Murphey, 2006) proposed an application of text mining in engineering diagnostics through automatic mapping of problem descriptions to the correct diagnostic categories. They proposed an automatic text categorization system which called DKETC (Diagnostic Knowledge Extraction and Text Classification). They conducted an experiment in order to study on a number of important issues relating to text document classification including term weighting schemes, latent semantic analysis (LSA), and similarity functions. The method was developed based on vector space model (VSM) which is one of the classic models in the information retrieval and text mining area. A text document categorization system used a training data set of 200,000 documents that describe 54 different categories of auto problem tested on a large test data collected from auto dealers.

Edwards et al. (Edwards, Zatorsky, & Nayak, 2008) used text mining techniques to analyze 12 years of data from dam pump station maintenance logs stored as free text in a spreadsheet application. The goal was to classify the data as scheduled maintenance or unscheduled repair jobs. The data is mined by calculating term weights to which clustering techniques are applied. Free text is converted to term weights and the singular value decomposition (SVD) is performed on term weights in order to create clusters. They applied classification models, namely, decision tree and neural network, to learn the cluster groups allowed the jobs to be identified in unseen data. They use SAS text mining software to perform their experiment.

Menon et al. (Menon, Tong, Sathiyakeerthi, Brombacher, & Leong, 2004) utilized text mining techniques in order to analyze textual data from product life cycle databases and extract unanticipated information and patterns quickly. Menon et al. (Menon, Tong, & Sathiyakeerthi, 2005) also tried to manipulate textual data mining in order to improve the quality and reliability of the product development process. The association rule mining is utilized and the plain text is used as an input to the text mining process. The methodology applied in two type of case studies, several service center databases, and the call center databases which were relevant to the product development process.

Romanowski and Nagi (Romanowski & Nagi, 2002) proposed a data mining method and graph theoretic which partially automated the process of creating generic bill of materials from legacy data. They applied unsupervised learning (clustering) methods to this process in order to cluster the BOMs from groups of products with similar patterns. The tree machining algorithms which are primarily for classification or determining cluster membership, are used in this experiment. The XML format is used for the text files of the BOM.

Romanowski and Nagi (Romanowski & Nagi, 2004) also proposed a data mining methodology in order to semi-automatically form a generic bill of material (GBOM), entities which represent diverse variants in product family, and expedite the search for similar designs and configuration of new variants. In this research, a novel method for generalizing parts and subassemblies using text mining is developed. They presented an algorithm for unifying similar BOM tree structures into a single GBOM as well as extracting design and configuration rules from the BOM data using association mining. The GBOMs were represented in Constrained XML format. The resulting GBOMs,



reduce the search space for retrieving similar previous designs and aid in configuring new variants. The paper concludes with a case study, using data from a manufacturer of nurse call devices, and identifies a new research direction for data mining motivated by the domains of engineering design and information.

Jiao et al. (J. R. Jiao, Zhang, Pokharel, & He, 2007) applied data mining techniques to identify generic routings from large amount of production information and process data which is available in a firm's legacy systems. Generic routing identification includes three consecutive stages, namely, routing similarity measure, routing clustering and routing unification. Text mining (i.e. Clustering) and tree matching techniques were applied to identify the textual and structural types of data underlying generic routings. A case study of mass customization production of vibration motors for mobile phones was conducted in order to exhibit the feasibility and potential of generic routing identification.

Jiao et al. (J. Jiao, Zhang, Zhang, & Pokharel, 2008) applied an association rule mining, using knowledge discovery from historical data in order to map the process and product variety. The mapping relationships are embodied in association rules, which can be deployed to support production planning of product families within existing production processes. For this application, a data mining tool, Magnum Opus was used. All relevant data are extracted from each transaction database and are input as a text file into Magnum Opus. The Magnum Opus provides five association metrics, namely, leverage, lift, strength, coverage and support. A case study of mass customization of vibration motors is presented to demonstrate how the association rule mining mechanism helps maintain the coherence between product and process variety. The performance of the association rule mining approach is further evaluated through sensitivity analysis.

Application of text mining has increased in recent years. But they are still limited and requires more attempts and progress. Researchers applied text mining in several production process and sub-processes. Several Methods such as association rule mining, Classification, Clustering and Tree Matching, as well as different techniques such as Support Vector Machine are being used in these areas of study. As expected, most of the knowledge extraction methods in the literature review are traditional text mining methods and using plain text as an input to their experiments. However, limited researches improved their methodology toward concept mining and concept based text mining methodology.

#### **4. Manufacturing Capability Thesaurus (MCT)**

The Manufacturing Capability Thesaurus (MC Thesaurus) is at the core of the proposed classification framework. It replaces the dictionary (bag-of-words) that is generated through machine learning in conventional text classification techniques (Bader, Berry, & Browne, 2008; Srivastava & Sahami, 2009). A formal thesaurus provides a high-quality collection of relevant terms and is free from the nuisances that are typically found in the bag-of-words. The MC Thesaurus captures the terms that directly or indirectly point to an aspect of manufacturing capability. Contract manufacturers may use terms and phrases such as precision machining, tool and die making, or build-to-order manufacturing to explicitly describe their technical capabilities, expertise, and services. Also, they may provide examples of parts they have produced in the past or industries and customers they have served in the past to advertise their capabilities indirectly. In the presence of a comprehensive thesaurus of manufacturing capability terms, it is possible to

readily translate each website into a vector model that is more amenable to quantitative analysis.

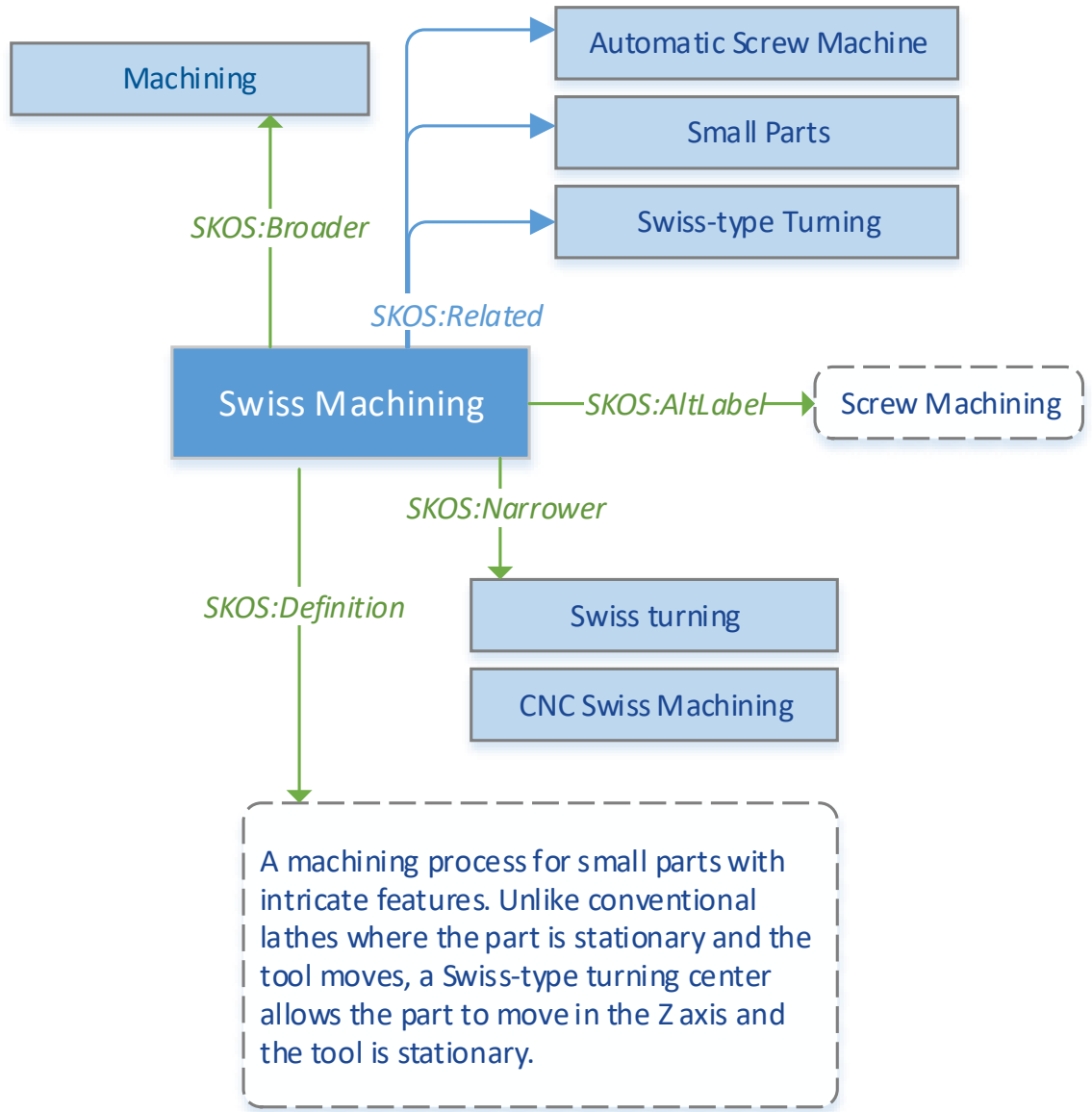
#### 4.1. SKOS

The thesaurus uses Simple Knowledge Organization System (SKOS) formalism. SKOS provides a structured framework for creating different types of controlled vocabulary such as thesauri, concept schemes, and taxonomies. SKOS thesauri are concept-based, as opposed to term-based, in nature. In a term-based thesaurus, terms are directly connected by lexical relationships whereas, in a concept-based thesaurus, semantic connection is at a concept level and terms are the lexical labels for the concepts and may or may not have lexical relationships established among themselves. For the purpose of this research, we define a concept as a unit of thought pointing to a *real entity* in the world. For example, “*the process by which three-dimensional objects are constructed by successively cutting material away from a solid block of material*” is the definition for a concept pointing to a manufacturing process that can be labeled by terms such as *machining*, *subtractive manufacturing*, and *material removal process*.

A SKOS thesaurus, like any other concept-based thesaurus, has a three-level structure (a) conceptual level, where concepts are identified and their interrelationships established; (b) terminological correspondence level, where terms are associated (preferred or alternative labels) to their respective concepts and (c) lexical level where lexical relationships (i.e., broader, narrower, synonym) are defined to interconnect the terms. The conceptual nature of SKOS is particularly useful in ontology development as it urges the developers to draw a distinction between terms and concepts and build a

sound conceptual understanding of the domain of discourse. Figure 1 shows the concept diagram for *Swiss Machining* process based on the SKOS semantics.

Each concept in SKOS has exactly one preferred label (*skos:prefLabel*) and can have multiple alternative labels (*skos:altLabel*). *Screw Machining* is the alternative label for *Swiss Machining* as it is used frequently for referring to the same concept. The broader concept of the *Swiss Machining* is *Machining*. *Swiss turning* and *CNC Swiss machining* are the narrower concepts; meaning that they are more specialized forms of *Swiss Machining*. The concepts that are related to *Swiss Machining* include *automatic screw machine*, *small part*, and *swiss type turning*. The reason *small part* is a *related* concept to *Swiss Machining* is that *Swiss Machining* is typically used for machining small parts with intricate features. In addition, each SKOS concept can have a definition provided in plain English or any other natural language. One major advantage of the SKOS thesauri is that they can be extended by community crowds and shared as linked data due to their open and standard syntax and semantics.



**Figure 1. The SKOS Concept Diagram for Swiss Machining Process**

#### **4.2. Thesaurus Development and Extension Process**

The MC Thesaurus is developed in a bottom-up manner through tagging the capability-related terms and phrases on the websites of manufacturing companies. Only frequently used terms are added to the thesaurus. The bottom-up approach enables capturing the informal terminology commonly used and well-understood in contract manufacturing industry. It should be noted that the developed thesaurus mainly addresses

the qualitative aspects of manufacturing capability. The quantitative aspects such as scrap rate, rejection percentages, tolerances and surface finish ranges are not included in the thesaurus.

A commercial thesaurus development system (Pool Party Thesaurus Management System, or PPT for short) was used for development of the MC Thesaurus. PPT supports bottom-up thesaurus creation through extracting relevant concepts from a text corpus that represents the terminology of the domain. The capability narratives of more than 380 contract manufacturing companies, mainly in precision machining industry, were imported into the PPT corpus to create a corpus with 382 documents. The terms with capability implications were collected through a semi-automated tagging mechanism. PPT provides two methods for tagging:

- 1) Directly tagging the relevant terms on the document text
- 2) Tagging recommended terms within the terms cloud generated for each document

During the tagging process, special attention was paid to the informal terms and phrases as well as the commonly used trade terms and acronyms that may or may not have scientific equivalents with textbook definitions. For example, *Turnkey Fabrication* is a frequently used phrase that refers to a project that the contractor undertakes the entire responsibility from design to manufacturing but this phrase cannot be found in any manufacturing handbook. A related concept to Turnkey Fabrication is *Full-Service Machine Shop* that was captured during tagging. Figure 2 shows the document frequency of some of the captured concepts based on the available set of document in the corpus.

The tagged terms are considered as *Candidate Concepts* before integration with the MC Thesaurus. The candidate concepts can be integrated with the thesaurus through identifying the appropriate *Broader Concept* for them.

The integrated concepts can be further described and formalized through the following steps:

1. Providing a textual definition of the concept
2. Providing alternative labels for the concept if possible
3. Linking the concept with the other related concept (both internal and external)

<b>Manufacturing Capability</b>		<b>Organizational Capability</b>	
<a href="#">Product</a>	282	<a href="#">Accreditation</a>	112
<a href="#">Assembly</a>	242	<a href="#">Small Business</a>	106
<a href="#">Part</a>	220	<a href="#">OEM</a>	38
<a href="#">Machining</a>	216	<a href="#">Continuous Improvement</a>	26
<a href="#">Material</a>	207	<a href="#">NADCAP</a>	9
<a href="#">Process</a>	194	<a href="#">Human Resources</a>	8
<a href="#">CNC Machining Center</a>	193	<a href="#">Information Technology</a>	4
<a href="#">CNC Machining</a>	177	<a href="#">dedicated people</a>	4
<a href="#">Fabrication</a>	149	<a href="#">Veteran Owned</a>	3
<a href="#">Stainless steel</a>	149	<a href="#">Highly Skilled Staff</a>	1
<b>General Capability</b>		<b>Industry Capability</b>	
<a href="#">Shipping</a>	144	<a href="#">Industry</a>	201
<a href="#">Precision Manufacturing</a>	80	<a href="#">Medical Industry</a>	103
<a href="#">Fabrication Service</a>	68	<a href="#">Defense Industry</a>	84
<a href="#">secondary operations</a>	53	<a href="#">Aerospace Industry</a>	39
<a href="#">Contract Manufacturing</a>	34	<a href="#">Oil And Gas Industry</a>	17
<a href="#">Custom Manufacturer</a>	27	<a href="#">Automotive Industry</a>	14
<a href="#">outsourcing</a>	24	<a href="#">Mining Industry</a>	8
<a href="#">General Capability</a>	22	<a href="#">Petrochemical</a>	8
<a href="#">Project Management</a>	20	<a href="#">Nuclear Power</a>	6
<a href="#">Custom Fabrication</a>	19	<a href="#">Military Industry</a>	5
<b>Engineering Capability</b>		<b>Quality Capability</b>	
<a href="#">Reverse Engineering</a>	40	<a href="#">Quality Control</a>	51
<a href="#">AutoCAD</a>	33	<a href="#">ISO-13485</a>	47
<a href="#">File Format</a>	32	<a href="#">ISO 9001</a>	34
<a href="#">SolidWorks</a>	31	<a href="#">AS9100</a>	28
<a href="#">CAD/CAM</a>	28	<a href="#">Quality Assurance</a>	28
<a href="#">IGES</a>	22	<a href="#">consistency</a>	17
<a href="#">CAD Design</a>	21	<a href="#">Final Inspection</a>	16
<a href="#">Computer-aided Manufacturing</a>	20	<a href="#">burr-free</a>	14
<a href="#">STEP</a>	20	<a href="#">Statistical Process Control</a>	13
<a href="#">Design For Manufacturing</a>	19	<a href="#">CMM inspection</a>	12

**Figure 2. Document Frequency of Some of the Concepts (Categorized Based on the Schema) in an Intermediate Stage of Thesaurus Development (Before Deleting Less Frequent Concepts)**

If an identified *top concept* contains only three or less narrower concepts, it is merged with other top concepts. Also, if a candidate concept appears in less than 5% of corpus documents, the concept is deleted from the thesaurus since it cannot be regarded as a recurring concept. At the time of preparing this document, MC Thesaurus contained about 760 preferred labels and 135 alternative labels.

### 4.3. Capability Model in MCT

The bottom-up term tagging and concept integration process described in the previous section resulted in creation of multiple categories of concepts (or *concept schemes*) as described below.

**Manufacturing Capability:** This concept scheme contains the concepts that directly point to the production capabilities of the company with respect to available processes, materials, equipment, geometries, production systems, and production capacity. Table 2 shows some of the top concepts and their corresponding narrower concepts under *Manufacturing Capability* concept scheme. Based on the document frequencies given in Figure 1, the concepts under this scheme are occurred in most documents in the corpus.

**Organizational Capability:** This concept scheme describes the capabilities related to the human resources, organization type, and industry-wide accreditations. Based on Figure 2, about one third of the documents in the corpus have used their *accreditations* as one indicator of organizational capability.



**Table 2: The Concepts under the Manufacturing Capability Concept Schema**

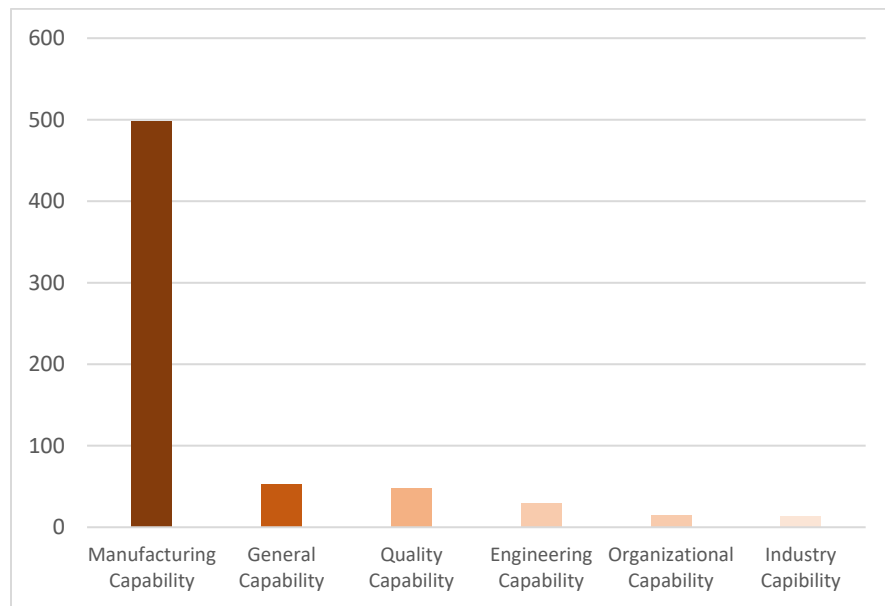
Top Concept	Representative Narrower Concepts
<b>Automation Capability</b>	Automatic work changer, In-line transfer, rotary transfer, tool positioner
<b>Material Testing Equipment</b>	Hardness tester, universal testing machine, fatigue testing machine
<b>Equipment Capability</b>	Automatic screw machine, CNC turning center, threading machine, vertical machining center, dual pallet machine.
<b>Geometric Capability</b>	Intricate Contours, Small diameter holes, Deep holes, Unusual shapes
<b>Material Capability</b>	Kovar, Capton, Nylon, Aluminum, exotic alloys
<b>Part Type</b>	Aluminum parts, Small parts, Welded assemblies, long tabular parts, structural frames, large machined parts, fabricated assembly
<b>Production Capacity</b>	Medium to large batches, full production batches, high mix of low volume, short runs, small prototype batches
<b>Production Support Capability</b>	ERP system, ERP/MRP system, Inventory Control

**Engineering Capability:** This concept scheme is a container for the concepts that point to the engineering capabilities of the company such as product design and development CAD/CAM, tool and fixture design, engineering analysis, reverse engineering and DFX. About 20% of the documents in the corpus contain some engineering capability concepts.

**Quality Capability:** The concepts under this schema are related to quality certifications, quality awards, quality control and inspection methods and tools and other term that are related to quality and inspection.

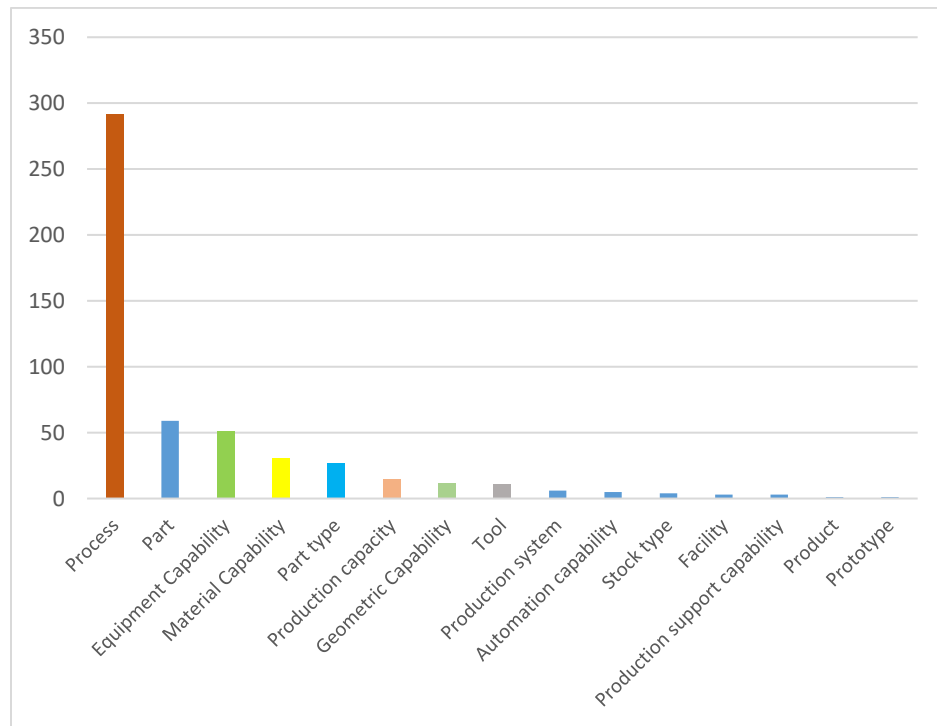
**Industry Capability:** This concept scheme contains the concepts that describe the market segments and industries the company has served such as defense, automotive, aerospace and so forth.

**General Capability:** The capabilities that cannot be categorized distinctly under the first four schemes are regarded as general capability. Build-to-order Service and Green Manufacturing are the examples of the concepts that are classified under General Capability scheme. Shipping is a general capability concept that has the highest document frequency, among other general capability concepts, in a corpus of 380 documents. It is highly likely that, as the thesaurus evolves in time, the general capability group is split into multiple well-defined groups. Total number of concepts under each concept scheme in the MC thesaurus is shown in Figure 3.



**Figure 3: Total Number of Concepts under Each Concept Scheme in the Manufacturing Capability Thesaurus**

As shown in this figure, manufacturing capability contains the highest number of concepts, followed by the general capability scheme, in the current version of the thesaurus. Figure 4 shows the number of concepts under different top concepts of the manufacturing capability scheme. This diagram indicates that the terms related to *processing*, *part*, and *equipment capabilities* are used more frequently in the capability narratives.



**Figure 4: Total Number of Concepts under Manufacturing Capability**

## 5. Manufacturer Classification Framework

The proposed framework for supplier classification has two distinctive features from a methodological perspective:

### 5.1. Bag-of-Concepts (BOC) Instead of Bag-of-Words (BOW) Method

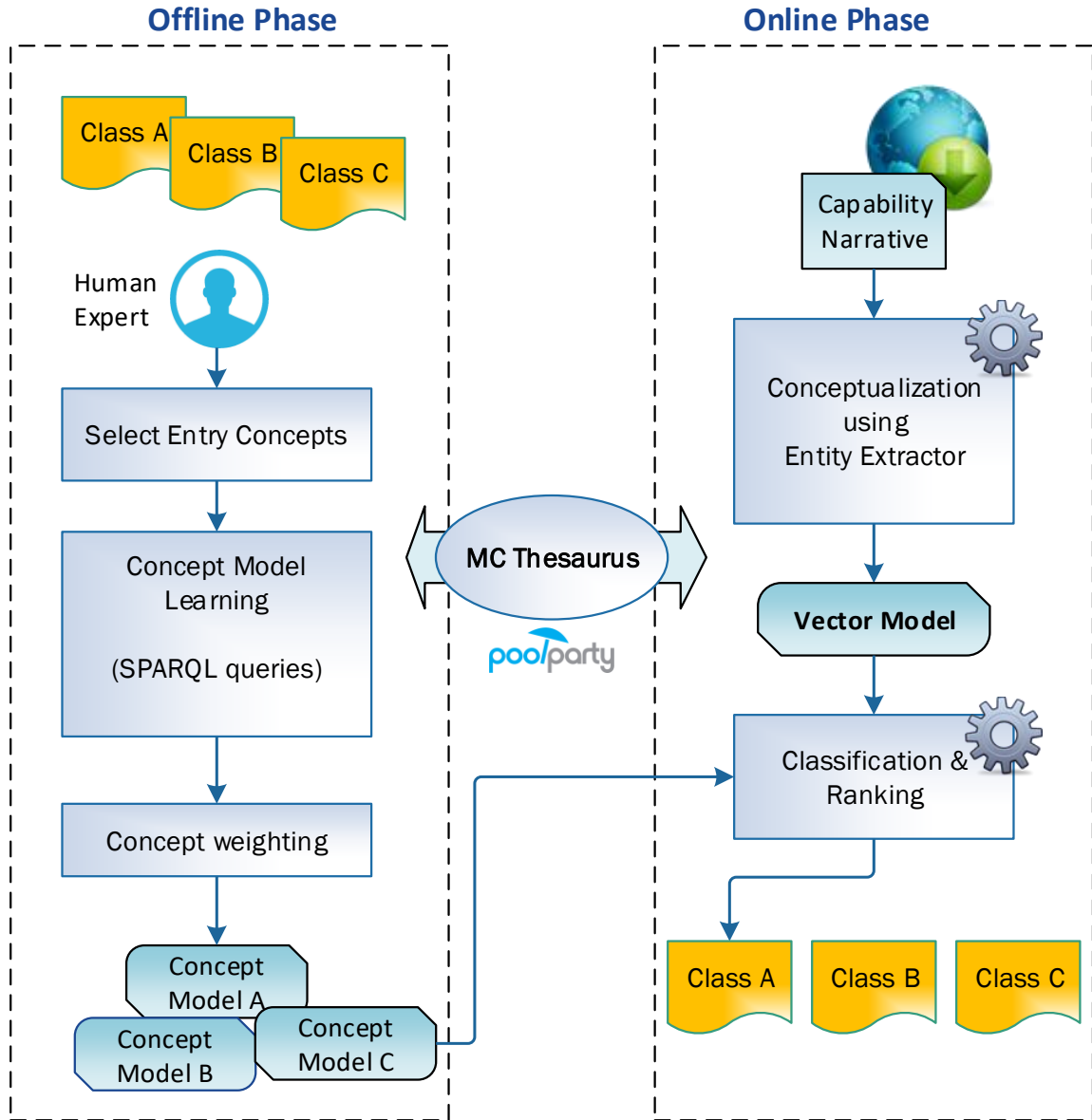
As discussed before, most of the well-known techniques for text classification are based on the *bag-of-words* method that represents documents as vectors of terms [17]. In the proposed framework, the *bag-of-concepts* method is being used instead in order to improve the semantic relevance of the results. In the BOC method, each concept is connected to other concepts through semantic relationships. MC Thesaurus provides the network of interrelated concepts that can guide the processes of selecting the representative concepts each class that form the *concept model* of the class of interest.

### 5.2. Naïve Bayes Method

The underlying mathematical model of the proposed classification framework is Naïve Bayes. Although Naïve Bayes is often outperformed by other more sophisticated classification methods, it is a popular baseline method since it is less computationally intensive and can produce meaningful output with a small training dataset.

As can be seen in Figure 5, the proposed framework is composed of two phases: (1) Concept Model Learning and (2) Test Document Classification. Concept Model Learning is the *offline phase* that results in generation of a set of representative concepts associated with each target class. Test document classification is the *online phase* that results in identification of the capability class for a given manufacturing company based on its capability narrative. The framework is implemented in R programming environment. Each phase is described in more details in the following sections. The concept model for a target class is a set of representative concepts, weighted based on their importance, which characterize the class with enough distinguishing power. The weighted concepts within the

concept model are treated as the *features* of the target class. Concept models can be learned through using a set of training documents or they can be extracted from the thesaurus through submitting relevant queries. In this work, the latter method is used.



**Figure 5: Proposed Manufacturer Classification Framework**

### 5.3. Concept Model Learning

**Entry Concepts:** The first step in building the concept model is to identify a few *entry concepts* ( $ec_i$ ) that can describe the capability class of interest. For example, if the target capability class is *the suppliers who are able to provide heat treating services*, then a possible entry concept set could be:

$$EC = \{ec_1=annealing, ec_2=hardening, \text{ and } ec_3=sintering\}$$

The queries that results in creation of the concept model for each class are built around the identified entry concepts of the class. The entry concepts are typically mid-level concepts that are highly connected to other concepts. In other terms, very abstract and high-level concepts or highly specific concepts should not be used as the entry concepts. It should be noted that the entry concept set for each target class is not unique and different entry concept sets can be used based on the classification strategy. The entry concepts are identified by the human expert.

Table 3 shows some example SPARQL queries that return broader, narrower, and related concepts for the Annealing concept. A set of SPARQL queries are needed to generate the complete concept model for each target class.

**Concept Weighting:** Different concepts have different levels of importance in the context of a specific capability class. For example, Gun Drilling is a more important concept for the Deep Hole Drilling capability class than the General Machining capability class. Figure 6 shows the weighting scheme that is used in this work. According to this scheme, the preferred and alternative labels of the entry concepts are weighted 9 and 5 respectively. The preferred or alternative labels of any related, narrower or broader

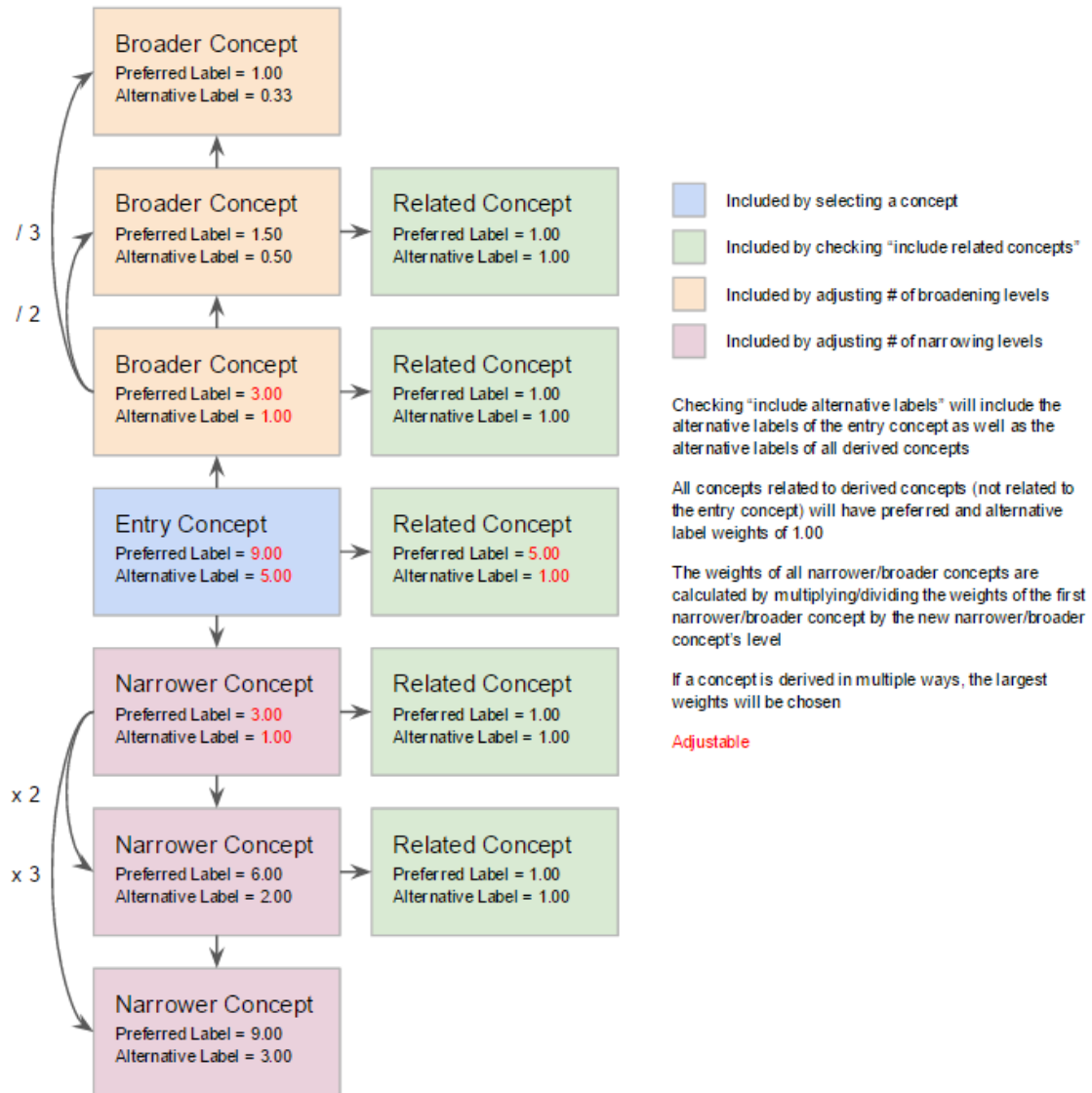
concept assumes lower weights based on the proposed scheme. The concepts within the thesaurus that are not returned by any query assume a weight of zero.

**Table 3: Concept Extraction via SPARQL Queries**

<b>Narrower concepts</b>	<b>SPARQL Query</b>	PREFIX skos:http://www.w3.org/2004/02/skos/core# SELECT ?label ?x WHERE {<http://infoneer.poolparty.biz/Processes/427> skos:narrower ?x. ?x skos:prefLabel ?label.}
	<b>Description</b>	This query returns narrower concepts of <i>annealing</i> .
	<b>Result</b>	"Recovery Annealing" "Recrystallization"
<b>Broader concepts</b>	<b>SPARQL Query</b>	PREFIX skos:http://www.w3.org/2004/02/skos/core# SELECT ?label ?m WHERE {<http://infoneer.poolparty.biz/Processes/427> skos:broader ?x . ?x skos:prefLabel ?label .}
	<b>Description</b>	This query returns broader concepts for <i>annealing</i> .
	<b>Result</b>	"Heat treating" "Hardening" "Sintering"
<b>Related concepts</b>	<b>SPARQL Query</b>	PREFIX skos:http://www.w3.org/2004/02/skos/core# SELECT ?label ?x WHERE {<http://infoneer.poolparty.biz/Processes/427> skos:related ?x . ?x skos:prefLabel ?label.}
	<b>Description</b>	This query returns related concepts of <i>annealing</i> .
	<b>Result</b>	"Casting"

**Concept Model:** The outcome of the offline phase is a concept model associated with a target class which is essentially a set of weighted concepts. The extracted concept model for capability class  $i$  can be represented as  $CM_i = ([c_1, w_1], [c_2, w_2], \dots, [c_m, w_m])$ , where  $m$  is the number of concept extracted for the capability class. The number of

generated concept models is equal to the number of target classes. Each concept model represents a row in the Document-Term Matrix (DTM) that will be used later in the classification step.



**Figure 6: Concept Weighting Schema**



## 5.4. Test Document Classification

In this phase, the objective is to convert suppliers' websites (documents) into concept vectors and classify them under the target capability classes via Naïve Bayes classifier. It is assumed that for each capability class, the associated concept model is already generated during the offline phase.

**Test Data Preparation:** For each supplier, a document is text document is created which contains the capability narrative directly copied from the company website. Some preprocessing such as removing numbers, stop words, and generic words is conducted on each document in this stage.

**Webpage Conceptualization:** Webpage conceptualization entails identifying a representative subset of concepts from the MC thesaurus that best describe the webpage. Through conceptualization, a webpage is translated into a *Concept Vector*. A custom-made tool, called *Entity Extractor Tool*, is developed and used for this step. The Entity Extractor Tool detects the thesaurus concepts that are appeared in a document through their preferred or alternative labels. It also calculates the frequency of occurrence for every detected concept. The tool receives plain text as the input and generates the concept vector associated with the text as a Comma-Separated Values (CSV) file. Different levels of abstraction can be used for the conceptualization process depending on the level of specificity of the target classes. For highly abstract and primitive classes, only top-level concepts are used for conceptualization whereas, for more specific and multi-dimensional classes, low-level concepts are also included in the vector model.

**Classification:** In this step, the capability classes associated with each supplier are identified. The training data (concept model for each capability class) and the test data (concept vector for each supplier) that are already converted into CSV format, are the inputs to the classifier. Finally, a classification algorithm, such as Random Forest or SVM, is used to assign a class label to each supplier in the dataset. The classification process is conducted in the R Studio environment.

## **6. Implementation and Experimental Validation**

In order to evaluate the efficiency and effectiveness of the proposed framework, an experiment is conducted based on two capability classes. The ultimate goal is evaluate the classifier based on information retrieval metrics such as precision, recall, and F-measure. Since the MC thesaurus is richer in terms of machining concepts, the target capability classes were formulated in the context of machining process capability. Examples of machining process capability include, high speed machining, high temperature alloy machining, heavy component machining, automotive machining, complex and difficult machining, and conventional machining. Each of these capability classes are expected to have a set of distinct features (concepts) that are different from other classes. For this experiment, *heavy component machining* and *difficult and complex machining* were selected as the target capability classes.

### **6.1. Class Definition**

Since the MC thesaurus in its current state is richer in terms of machining concepts, the target capability classes were defined based machining process capability. Examples of machining capability classes include, high speed machining, high

temperature alloy machining, heavy component machining, automotive parts machining, complex and difficult machining, and conventional machining. Each of these capability classes often have a set of distinct features (concepts) that uniquely characterizes the class. For this experiment, heavy component machining and difficult and complex machining were selected as the target capability classes. Heavy machining capability is a part quality capability (second-order capability) and is defined as the ability to machine large and heavy parts weighting up to several tons. Complex machining capability refers to the ability to machine parts with intricate and geometrically complex features. While it is quite likely to encounter companies that have expertise in both areas, typically it is difficult for small-to-medium sized companies to obtain both capabilities due the need for different types of specialized production machinery and transportation equipment. Therefore, it was expected that the classification process would result in fairly distinct classes of suppliers with some partial overlap.

**Table 4: Entry Concepts for Target Classes**

Heavy Component Machining	Difficult and Complex Machining
“heavy component” “large part”	“complex machining”
“vertical machining center”	“difficult machining”
“deep hole machining”	“live tooling”
“heavy Machining”	“complex precision part”
“large CNC machining”	“multi-axis capabilities”

## 6.2. Generating Concept Models for Heavy and Complex Machining Classes

Based on the proposed framework, the first step of the offline phase entails building the concept model for each class through identifying the entry concepts and then generating the concept model by submitting appropriate SPARQL queries that are formulated around the entry concepts. The entry concepts for both target classes are listed in Table 4. The entry concepts for the classes of interest were determined by the domain expert. Once the entry concepts were selected, multiple queries were submitted to the MC thesaurus. The concept model is then populated by the related, narrower, and border concepts of each entry concept. The concepts were then weighted based on the provided weighting schema. For example, Vertical Boring, Gantry Mill, and Bridge-type CMM are examples of the concepts that are semantically and contextually related to Heavy Machining concept. Therefore, they were returned by the submitted queries and included in the concept model for Heavy Machining class. To facilitate the process of generating concept models for each class, the Concept Model Builder (CMB) module was embedded in the SKOS Tool as shown in Figure 7.

Select concept #1:

Live Tooling

☒ Include related concepts

☒ Include alternative labels

☐ Include top-level concepts

Narrowing levels: 2

Broadening levels: 1

+ Add concept

Submit

**Figure 7. Concept Model Builder Function**

Using this CMB module, the user can select the entry concepts for the target class and specify the type and depth of the SPARQL queries that need to be submitted to the thesaurus. SPARQL queries can become quite complex which makes manual formulation of the queries very tedious and time-consuming. It also requires familiarity with SPARQL syntax. The CMB function eliminates the need for manual query formulation and submission.

The CMB returned 39 concepts for Heavy Component Machining and 44 concepts for Difficult and Complex Machining. At the time the queries were submitted for this experiment, the MC Thesaurus contained 762 preferred labels and 208 alternative labels. Each returned concept is weighted according to the weighting scheme given in Figure 6. The weight of the concept denotes the strength of the semantic relationship between the concept and the target capability class. Abstract and generic concepts, such as Machining, have lower weighting since they are not regarded as major discriminating features when classifying a supplier. The final concept model is exported as a CSV file to be used as the input to the classifier. Table 5 shows some of the members of complex machining and heavy machining concept models together with their weightings.

**Table 5: Some of the Members of Complex Machining and Heavy Machining  
Concept Models**

Concepts	Weightings	
bridge type CMM	3	1
Deep Hole	3	1
Deep Hole Machining	5	1
Heavy Lifting Equipment	5	1
Heavy payload	5	1
Large and complex part	9	1
Large CNC Vertical Machining	5	1
Large working envelope	5	1
vertical machining center	9	1
Vertical Turning Lathe	5	1
7-Axis Machining	1	5
Complex CNC Machining	1	5
complex dimensional shapes	1	5
Complex Machining	1	9
complex precision parts	1	5
Complex turning and milling	1	3
Dicult machining	1	9
multi-axis capabilities	1	9
Multi-axis Complex Machining	1	3
Live Tooling	1	5
Class	Heavy Machining	Complex Machining

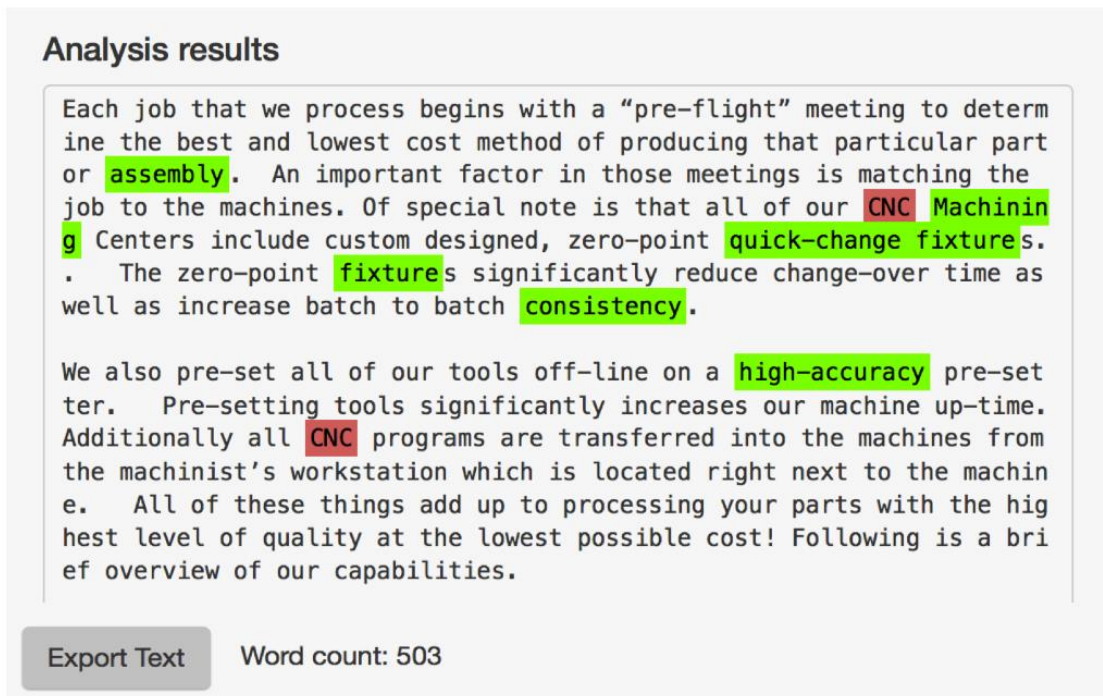
### 6.3. Test Data Preparation

For this classification experiment, 260 suppliers in contract machining industry were selected and their capability narratives were imported as plain text files. The participating suppliers were selected from Thomas Net repository. Thomas Net is a web-based sourcing portal that contains multiple categories of suppliers including heavy machining and complex machining suppliers. Equal number of suppliers from each class were selected from Thomas Net categories. The label (class) of each supplier was determined by its Thomas Net category. In order to reduce the noise induced by the irrelevant words, only the text from the pages that directly describe the manufacturing capabilities and services of the company was selected.

The screenshot shows a web interface titled "Upload text". It features two radio buttons: "Manual" (unselected) and "URL (case sensitive)" (selected). Below these is a large empty text input field. Underneath the input field is a label "Select thesaurus:" followed by a dropdown menu currently displaying "MCT". Below the dropdown are four options with checkboxes: "Upload new thesaurus" (with an upload icon), "Delete all thesauruses" (with a delete icon), "Include zero-occurrence concepts" (unchecked), and "Include top-level concepts" (unchecked). There is also a checked checkbox for "Show URL preview page". Below these options is a slider for "URL depth:" set to "1". At the bottom left is a grey button labeled "Analyze".

**Figure 8. The User Interface for Extracting Capability Text**

Test data preparation process is facilitated by the Entity Extractor (EE) module of the SKOS Tool. The EE module receives the URL of the supplier and, after some preprocessing, generates a single text file that aggregates the capability text collected from multiples web pages under the same URL. The preview option enables the user to only pick the relevant pages and exclude the pages that do not contain capability data. Alternatively, the user can directly insert the text in the provided text box or import the text as a CSV file. The EE interface used for extracting text from suppliers URLs is shown in Figure 8. Once the URL (or the plain text or the CSV file) is submitted for analysis, the EE module generates the concept vector of the input text.



**Figure 9. Sample Capability Narrative Tagged by MC Thesaurus Concepts**

Figure 9 shows the sample capability narrative which is tagged by manufacturing capability thesaurus concepts. Figure 10 shows the extracted concepts and frequencies of the sample capability narrative.



Sort table:

☐ Alphabetical ☒ Occurrences

Concept (preferred label)	Occurrences
CNC Machining Center	15
Machining	7
vertical machining center	4
Computer-aided Manufacturing	4
Fixture	3
SolidWorks	2
quick-change fixture	2
STEP	1
Quality Assurance	1
high-speed spindle	1

**Figure 10. Extracted Concepts and Frequencies of the Sample Capability Narrative**

#### 6.4. Test Data Conceptualization

Data conceptualization involves creating the concept vector corresponding to each document. As mentioned before, each document contains the aggregated capability text collected from the website of a given supplier. The concept vector for each document is generated automatically using the Entity Extractor module of the SKOS Tool. The EE module generates a list of detected concepts, together with their frequencies, as a CSV file. Once the concept vectors for all suppliers are generated, they are combined to form a single matrix. The resulting document-term matrix (DTM) is partially shown in Table 6.

**Table 6. Concept Vectors for 10 Suppliers Belonging to Heavy Machining and  
Complex Machining Classes**

Concepts	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
<b>bridge type CMM</b>	0	3	7	0	1	0	0	0	4	0
<b>Deep Hole</b>	7	1	0	3	0	0	1	0	1	0
<b>Deep Hole Machining</b>	1	0	6	0	9	0	0	7	1	0
<b>Heavy Lifting Equipment</b>	0	8	0	1	3	0	1	0	0	0
<b>Heavy payload</b>	7	0	2	0	0	1	0	0	0	0
<b>Large and complex part</b>	2	0	1	2	0	0	0	0	6	0
<b>Large CNC Vertical Machining</b>	2	0	6	4	0	0	2	0	0	0
<b>Large working envelope</b>	0	3	0	1	0	0	0	0	0	0
<b>vertical machining center</b>	7	0	2	0	4	2	0	0	0	4
<b>Vertical Turning Lathe</b>	0	8	1	0	2	0	0	0	0	0
<b>7-Axis Machining</b>	0	0	1	0	0	1	0	8	3	0
<b>Complex CNC Machining</b>	1	3	9	0	0	0	0	0	2	1
<b>complex dimensional shapes</b>	0	0	0	0	0	1	3	0	0	2
<b>Complex Machining</b>	8	4	2	0	1	3	6	1	0	3
<b>complex precision parts</b>	0	0	0	0	0	0	0	3	0	0
<b>Complex turning and milling</b>	0	0	0	0	0	0	0	3	1	7
<b>Difficult machining</b>	0	0	0	0	0	3	6	1	2	5
<b>multi-axis capabilities</b>	0	0	0	0	1	2	0	0	2	3
<b>Multi-axis Complex Machining</b>	1	0	0	2	0	0	0	0	2	0
<b>Live Tooling</b>	0	2	0	0	0	2	0	4	0	0
<b>Class</b>	HM	HM	HM	HM	HM	CM	CM	CM	CM	CM

In this matrix, the rows indicate the concepts and the columns indicate the suppliers. The full matrix has 260 columns and more than 50 rows since the concept models of both classes are merged together in this matrix. The cell values indicate the frequency of occurrence of concepts in suppliers' capability narratives. To investigate the impact of concept weightings, another version of DTM was created in which the cell values were determined by multiplying the frequency of each concept by its weight. The classification is then conducted under two scenarios:

(1) Scenario one: Using the DTM matrix without weightings

(2) Scenario two: Using DTM matrix with weightings.

The next step is to classify the suppliers using different classification techniques.

### **6.5. Performance Evaluation**

To evaluate the accuracy of the classifier three metrics were used, namely, precision, recall, and F-measure. Precision is the fraction of suppliers that are correctly classified for a given class while recall is the ratio of number suppliers that are correctly classified over the total number of suppliers that are known to belong to a given class. F-measure is a score which is calculated based on combination of precision and recall.

The F-measure is used to measure the model accuracy. In order to calculate these measures for the target classes, the Equations 5.1 to 5.3 were used. Table 7 shows the interpretation of Positive and Negative results for heavy component machining class.

**Table 7: Data Required for Calculation of Precision, Recall, and F-measure for Heavy Machining**

<b>True Positive (TP)</b> Number of heavy machining suppliers which are classified as heavy machining class.	<b>False Negative (FN)</b> Number of heavy machining suppliers which are not classified as heavy machining class.	<b>TP + FN</b>
<b>False Positive (FP)</b> Number of non-heavy machining suppliers which are classified as heavy machining class.	<b>True Negative (TN)</b> Number of non-heavy machining suppliers which are not classified as a heavy machining class.	<b>FP + TN</b>
<b>TP + FP</b>	<b>FN + TN</b>	<b>N</b>

$$Precision = \frac{TP}{TP + FP} \quad (5.1)$$

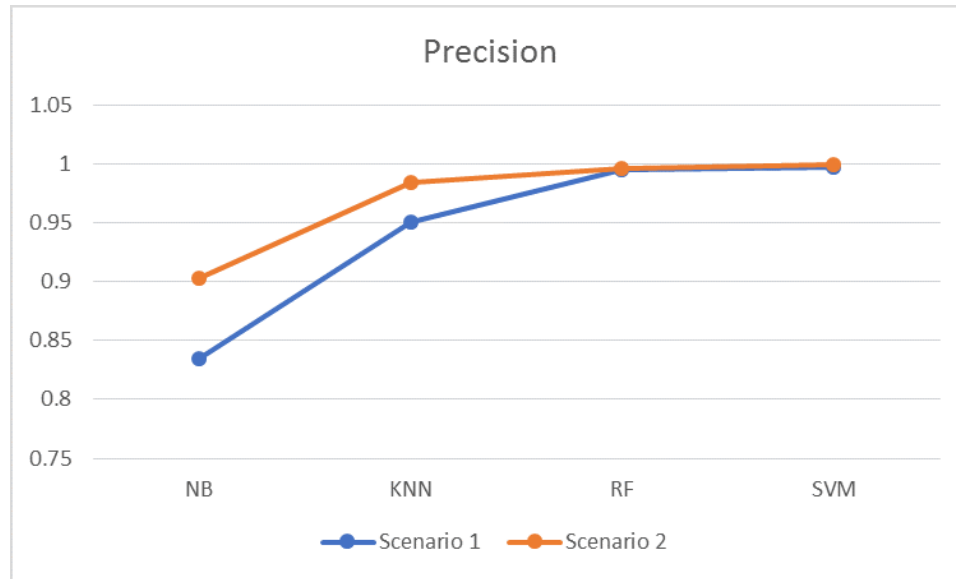
$$Recall = \frac{TP}{TP + FN} \quad (5.2)$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5.3)$$

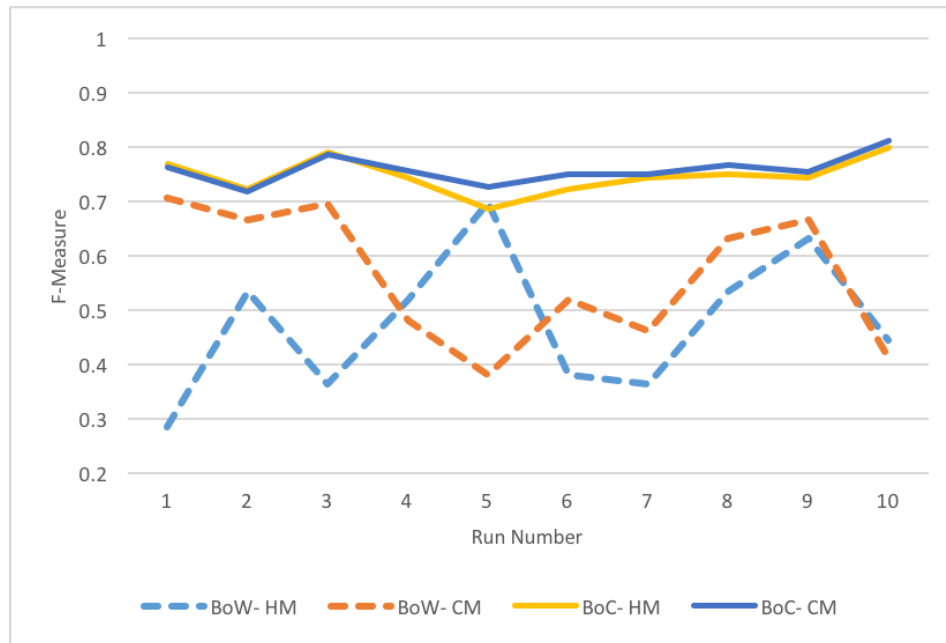
## 6.6. Results

Four classification techniques, namely, Nave Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM), were used. For all techniques, about 75% of the data was randomly selected as training data and the rest was regarded as the test data. To eliminate the bias caused by the specific choice of training data, the classification was run for 10 times per technique for both scenarios. Table 8 shows the detail result of the experiments. The results indicate that the SVM technique has the best performance among the used techniques with an average precision of 99%. As expected, Naive Bayes method was the low performer as was only used to provide a baseline for comparison. Another important observation was the overall precision

improves when concept weightings are applied (scenario two). Also, Figure 12 shows that BOC method has higher accuracy (i.e. F-Measure) than BOW method.



**Figure 11. The Precision of the Text Classification**



**Figure 12. Bag of Concepts VS Bag of Words**

**Table 8. Detail Result of the Two Scenarios and Four Classification Techniques**

Run	Naïve Bayes			K-Nearest Neighbor			Random Forest			Support Vector Machine		
	P <sup>2</sup>	R <sup>3</sup>	F <sup>4</sup>	P	R	F	P	R	F	P	R	F
	Scenario 1: Without Weightings											
1	0.810	0.864	0.836	0.969	0.967	0.968	0.982	0.981	0.982	1.000	1.000	1.000
2	0.824	0.871	0.847	0.955	0.950	0.952	1.000	1.000	1.000	0.994	0.994	0.994
3	0.797	0.658	0.721	0.969	0.967	0.968	1.000	1.000	1.000	1.000	1.000	1.000
4	0.837	0.822	0.829	0.941	0.944	0.943	0.988	0.988	0.988	1.000	1.000	1.000
5	0.873	0.904	0.888	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6	0.865	0.902	0.883	0.919	0.923	0.921	1.000	1.000	1.000	1.000	1.000	1.000
7	0.808	0.691	0.745	0.975	0.971	0.973	1.000	1.000	1.000	0.978	0.975	0.976
8	0.818	0.761	0.789	0.876	0.873	0.875	1.000	1.000	1.000	1.000	1.000	1.000
9	0.841	0.778	0.808	0.899	0.887	0.893	0.975	0.975	0.975	1.000	1.000	1.000
10	0.875	0.833	0.854	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Average	0.835	0.808	0.820	0.950	0.948	0.949	0.994	0.994	0.994	0.997	0.997	0.997
Scenario 2: With Weightings												
1	0.875	0.833	0.854	0.955	0.950	0.952	1.000	1.000	1.000	1.000	1.000	1.000
2	0.917	0.927	0.922	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	0.937	0.939	0.938	1.000	1.000	1.000	0.988	0.988	0.988	1.000	1.000	1.000
4	0.905	0.910	0.908	0.984	0.976	0.980	1.000	1.000	1.000	1.000	1.000	1.000
5	0.897	0.907	0.902	0.960	0.960	0.960	0.976	0.988	0.982	1.000	1.000	1.000
6	0.916	0.923	0.919	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7	0.919	0.931	0.925	0.972	0.972	0.972	1.000	1.000	1.000	0.994	0.988	0.991
8	0.851	0.833	0.842	0.968	0.968	0.968	1.000	1.000	1.000	1.000	1.000	1.000
9	0.927	0.935	0.931	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
10	0.890	0.895	0.893	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Average	0.903	0.903	0.903	0.984	0.983	0.983	0.996	0.998	0.997	0.999	0.999	0.999

<sup>2</sup> Precision

<sup>3</sup> Recall

<sup>4</sup> F-Measure

**Table 9. T-Test to Compare the Result of Scenario One and Two for Different Classification Techniques**

Classification Technique	F-Measure Average		P-value	
	Scenario One	Scenario Two	One-tailed	Two-tailed
Naïve Bayes	0.820	0.903	0.000314	0.000629
K-Nearest Neighbor	0.949	0.983	0.016173	0.032346
Random Forest	0.994	0.997	0.248326	0.496652
Support Vector Machine	0.997	0.999	0.212359	0.424719

In order to measure the influence of the scenarios in accuracy of the different classification techniques, statistical analysis is investigated. Since the number of experiment runs are less than 25, one-tailed and two-tailed T-test with the significant level of 0.05 is conducted. Data is used from Table 8 in order to conduct this experiment. It is shown in Table 9 that P-value of both one-tailed and two-tailed experiment in the Naïve Bayes and K-Nearest Neighbor classification techniques are less than 0.05 which means that scenario two has a significant impact on improving the accuracy of these classification techniques. For Random Forest and Support Vector Machine, the improvement is not significant since the accuracy in scenario one is already high.

**Table 10. T-Test to Compare the Result of Bag of Words and Bag of Concepts Approach for Naïve Bayes Classification Techniques**

Classification Technique	F-Measure Average		P-value	
	Bag of Words	Bag of Concepts	One-tailed	Two-tailed
Naïve Bayes	0.5535	0.7584	2.48E-08	4.97E-08

Another T-test with same level of significance is also conducted in order to compare the accuracy of Naïve Bayes classifier in BOW and BOC approaches. It is concluded that the accuracy (i.e. F-measure) in Bag of Concept method is significantly higher than the accuracy in Bag of Words method since the P-value is very small.

## 7. Conclusion

Capability-based classification is a necessary first step for accurate supplier evaluation in decentralized scenarios. In this paper, a novel framework for capability-based supplier classification based on unstructured capability data is proposed. The proposed framework uses a concept-based method and is supported by a SKOS-based thesaurus referred to as Manufacturing Capability Thesaurus. The thesaurus encodes the domain knowledge and provides a semantically connected network of capability concepts. The MC Thesaurus is developed in a bottom-up fashion through tagging the key terms on suppliers' websites. The novel feature of the MC Thesaurus is that it contains the informal terms typically used in contract manufacturing industry. The MC Thesaurus can be extended and validated by domain experts in a collaborative fashion using cloud-based tools. Therefore, it can serve as a trustable source of manufacturing capability concepts. Furthermore, since it is based on SKOS, it can be linked to various open-source datasets on Linked Open Data (LOD) and reuse the existing concept models, thus enabling continuous and dynamic evolution and extension. Additionally, because SKOS thesauri are machine-understandable, the MC thesaurus can be integrated seamlessly with different semantic solutions that can support supply chain decisions. The crowdsourced extension and validation of MC Thesaurus significantly reduces the cost of evolution and validation. The proposed document classification framework is generic enough that can be applied to other areas such as customer review analysis or maintenance report classification.



## Chapter III

### MANUFACTURING SUPPLIER CLUSTERING

#### 1. Introduction

Following the previous research work (Sabbagh & Ameri, 2017; Sabbagh, Ameri, & Yoder, 2018) on classification of manufacturing suppliers based on their unstructured capability data available on their websites, the next step is the investigation of the associations and relations between different terms in this data. Only a small number of the suppliers have limited their activities into one specific manufacturing service. Most of the manufacturing suppliers offer several types of services, processes, and products. When comparing two manufacturing suppliers given that Supplier I offers only one specific type of services such as *casting* and Supplier II offers 10 types of services including *casting*. When dealing with large number of manufacturing suppliers and several types of categories based on their services, it is not fair to label a supplier with a single category when it offers several services. On the other hand, it is not fair for the manufacturing supplier that offers only one service in a high quality and quantity to be equalize to a supplier that offers several services but in limited quality and quantity.

The data available in manufacturing suppliers' website is unstructured. In order to deal with the unstructured data and extract information out of it, there are several supervised and unsupervised learning techniques. In the supervised learning techniques, documents should be labeled so that they could serve as training data. The test data documents could be labeled in order to measure the method's accuracy. It could be unlabeled as well to explore these documents and realize the appropriate categories they

are associated with. On the other hand, if all of the documents are unlabeled, the unsupervised learning algorithms are more appropriate. Unsupervised learning is a type of machine learning algorithm, which could describe hidden structure from unlabeled data (Hastie, Tibshirani, & Friedman, 2009). The most popular unsupervised learning method is cluster analysis (Jain et al., 1999), which is used for exploratory data analysis to find the hidden patterns or grouping in data.

Moreover, when using supervised learning, in most cases the nature of the data is known and the goal is to expand the knowledge in specific well-known area. But when conducting unsupervised learning, the goal is to find novel concept, novel characteristics, novel fields, etc. The purpose of this research is to help users efficiently categorize existing unstructured data and extract new information from manufacturing suppliers' website. This article proposes a model based on clustering and topic modeling methods in order to facilitate online search and organization of manufacturing capability terminology as well as extraction of novel patterns in manufacturing corpora. This could help experts discover novel characteristics and categories in manufacturing industry. Two major problems associated with unsupervised learning research are how to automatically find new and important terms from manufacturing supplier's website and how to identify, classify, and categorize the new terms, which are detected from supplier's website. These two questions will be addressed separately.

## **2. Related Work in Unsupervised Learning**

Data mining in both supervised and unsupervised forms has been applied in diverse areas ranging from publishing and media, banks, insurance and financial markets

to political institutions (Gupta, Lehal, & others, 2009). In the manufacturing domain, however, it is relatively a new approach (Barazandeh et al., 2017; Bastani, Barazandeh, & Kong, 2018). In this section, some of the applications of unsupervised learning methods are briefly discussed.

Xue and Dong (Xue & Dong, 1997) proposed a feature modeling system using two types of features, design features and manufacturing features, for modeling these two product life-cycle aspects. Design features which are represented as mechanical components and mechanisms, are used for modeling design candidates to satisfy design functions. A design feature coding system was developed based on the analysis of design functions. A fuzzy pattern clustering algorithm was employed to organize the large design feature library into hierarchical feature groups. Required design features are identified using graph-based search. Finally, a group-technology-like approach was introduced to organize components into groups according to their manufacturing feature codes using a fuzzy clustering algorithm. Production operations are optimized by a special optimization module. The two coding systems have been implemented in a feature-based, integrated concurrent design system for generating design candidates and planning production processes.

Torkul et al. (Torkul, Cedimoglu, & Geyik, 2006) studied a fuzzy logic approach in design of part families and machine cells simultaneously in order to compare manufacturing cell design which made of fuzzy clustering algorithm (Fuzzy C-Means) with the crisp methods and investigate the applicability of these techniques. It has been seen from the result of the study, fuzzy clustering solutions may be efficient than the crisp method for the selected data sets.

Chen and Lee (Y. Chen & Lee, 2011) proposed a data mining method to search historical alarm logs for the correlations that can represent causal relationships in order to increase the analytical capability of the alarm analysis. A hierarchical clustering method was used to carry out the correlation pattern search. Moreover, the similarity function of the method was designed to identify certain pre-defined correlation patterns. This method was validated in a vertical turning machine center alarm system application. The method could discover a large number of alarm correlations, which were usually neglected by operators, and manage the alarms in the way that clarify process disturbance and enabled rapid root cause analysis.

Chen and LeClair (C. L. P. Chen & LeClair, 1994) presented an approach in order to integrate the design and manufacturing knowledge. The method utilized a feature-based design environment and an unsupervised learning algorithm to categorize features into a setup for machining. The proposed algorithm and architecture incorporated multiple objective functions into setup generation. Intersecting and nonintersecting features within a setup were identified and classified using an associative memory. A discover-and-merge algorithm merged the tool graphs of features into a new tool graph. An optimal-tool-sequence algorithm was introduced to find the best sequence across the features in a setup.

Zhai et al. (Zhai, Liu, Xu, & Jia, 2011) first extended a popular topic modeling method, called Latent Dirichlet Allocation (LDA), with the ability to process large scale constraints and proposed two novel methods to extract two types of constraints automatically. The aim of the research was to produce a summary of opinions based on product features and attributes in opinion mining of product reviews. They added some

pre-existing knowledge to the topic modeling in the form of automatically extracted constraints in order to better find the groupings. Finally, the resulting constrained-LDA and the extracted constraints were applied to group product features. Experiments showed that constrained-LDA outperformed the original LDA and the latest multilevel latent semantic association (mLSA) method by a large margin.

Tan et al. (Tan et al., 2014) proposed a Latent Dirichlet Allocation (LDA) based model, Foreground and Background LDA (FB-LDA), in order to distill foreground topics and filter out longstanding background topics. These foreground topics could give potential interpretations of the sentiment variations. To further enhance the readability of the mined reasons, they selected the most representative tweets for foreground topics and developed another generative model called Reason Candidate and Background LDA (RCB-LDA) to rank them with respect to their “popularity” within the variation period. Experimental results demonstrated that the methods could effectively find foreground topics and rank reason candidates. The proposed models could also be applied to other tasks such as finding topic differences between two sets of documents.

Yazdizadeh and Ameri (Shotorbani et al., 2016) proposed an approach in order to facilitate the search and organize the textual documents and also extract the thematic patterns in manufacturing corpora using document clustering and topic modeling techniques. The proposed method adopted K-means and Latent Dirichlet Allocation (LDA) algorithms for document clustering and topic modeling, respectively. Through experimental validation, it was shown that topic modeling, in conjunction with document clustering, facilitated the automated annotation and classification of manufacturing

webpages as well as extraction of useful patterns, thus improving the intelligence of supplier discovery and knowledge acquisition tools.

Malakooti and Yang (Malakooti & Yang, 1995) developed an unsupervised learning clustering neural network approach in order to solve the machine-part group formation problem. They also developed a neural network clustering system in order to cluster the 0-1 matrix into diagonal blocks. The algorithm considered the lower and upper bounds on the number of machines in each cell. The computational results were comparably close to those from well-known rank order clustering and directive clustering methods.

### **3. Overview of the Framework**

Similar to most text mining techniques, the very first step of the proposed framework is preprocessing the extracted text from the web-based resources. The contents of the websites of manufacturing companies are very heterogeneous in terms of syntax and semantics and contain a lot of tribal knowledge, acronyms, and trade jargons. Therefore, it is necessary to apply several preprocessing steps such as removing numbers, symbols, punctuation, hyphens, and stop words to the extracted text give some uniformity to the input documents. The next step is reducing the dimensionality of the dataset, which means detecting the terms with capability significance and removing the less significant terms. For this purpose, the Latent Semantic Analysis (LSA) technique is used in combination with N-gram analysis. The output of LSA, which is a normalized document-term matrix (DTM), is used as the input to the K-means clustering method. In order to examine the accuracy of the generated clusters based on precision, recall, and F-measure,

a dataset composed of two *known* clusters is analyzed. Once the accuracy of the clustering method based on this dataset is validated, the clustering method is applied on a dataset of unseen manufacturing suppliers to build new clusters of similar suppliers. Topic modeling technique is then used for discovering the main theme of each cluster. The details of the aforementioned steps are described in the following sections.

### **3.1. Extracting the Raw Text from Manufacturing Supplier's Website**

The raw text which is available on manufacturing supplier's website is imported into separate text files. A web-based tool, developed internally, was used for crawling the websites and importing their contents into plain text files without any images or tables. The text is extracted only from the pages that contain capability data.

### **3.2. Preprocessing**

In this step, the plain text documents are converted into the csv format. For this purpose, the text files are collected by *read. Text* function in the R and integrated as a single csv file. The punctuations, hyphens, numbers, and symbols are removed and all the letters are converted to lowercase. Several stop words such as “can”, “also”, “yet”, “offer”, “capable”, “need”, “time”, “contact”, “available”, etc. are also applied to the documents. The stemming function is avoided because it would potentially result in semantic degradation of the data.

### **3.3. Extracting N-grams after Preprocessing**

The N-gram analysis is used to capture all possible combination of terms, which appeared together without any other terms in between. N-gram is a set of co-occurring

words extracted from a given text and  $N$  is an integer number that determines the length of  $N$ -gram. For example, for the sentence “*we can machine different exotic materials*”, if  $N=3$ , the possible 3-grams (also known as trigrams) are:

- *we can machine*
- *can machine different*
- *machine different exotic*
- *different exotic materials*

Obviously not all of the generated  $N$ -grams are meaningful and there is a need to filter them based on their discriminating power before using them for cluster analysis. The  $N$ -gram function in R investigates all two, three, or more combinations of the terms.  $N$ -grams augment the document-term frequency matrices. This often leads to increased performance (e.g. accuracy) for machine learning models trained with more than just unigrams (i.e. single terms). The extracted  $N$ -grams (unigrams, bigrams, and trigrams) are added as the columns of a document-term matrix. The body of the matrix represent the frequencies of the detected  $N$ -grams in each document (rows of the matrix). However, the frequency of an  $N$ -gram should not be used directly as the measure of importance or informativeness of the  $N$ -gram since the size of documents as well as the level of commonality of the terms could wrongfully inflate the importance of the  $N$ -gram. Therefore, some normalization is required.

### **3.4. TF-IDF Normalization on Term Frequency**

Term Frequency-Inverse Document Frequency (TF-IDF) is a powerful function for enhancing the information and signal contained within the document-frequency



matrix (Ramos & others, 2003). Specifically, the mathematics behind TF-IDF accomplish the following goals:

- i. When comparing multiple documents, the longer documents would have higher individual term counts than the shorter documents. Since the term frequencies are important in classification and clustering process, the TF function is used in order to eliminate the effect of document's length on data analysis process:

$$TF(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}} \quad (3.1)$$

- ii. It is obvious that the term, which appears in most of the documents, has little power to predict the class of the document. The IDF function weighs down the frequent terms in the corpus while scaling up the rare ones:

$$IDF(t) = \log_{10}\left(\frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}\right) \quad (3.2)$$

- iii. The multiplication of the TF and the IDF functions will consider both frequencies with regards to the document's length as well as the frequencies of the terms in all documents in the corpus.

### 3.5. Identifying the Useful N-grams

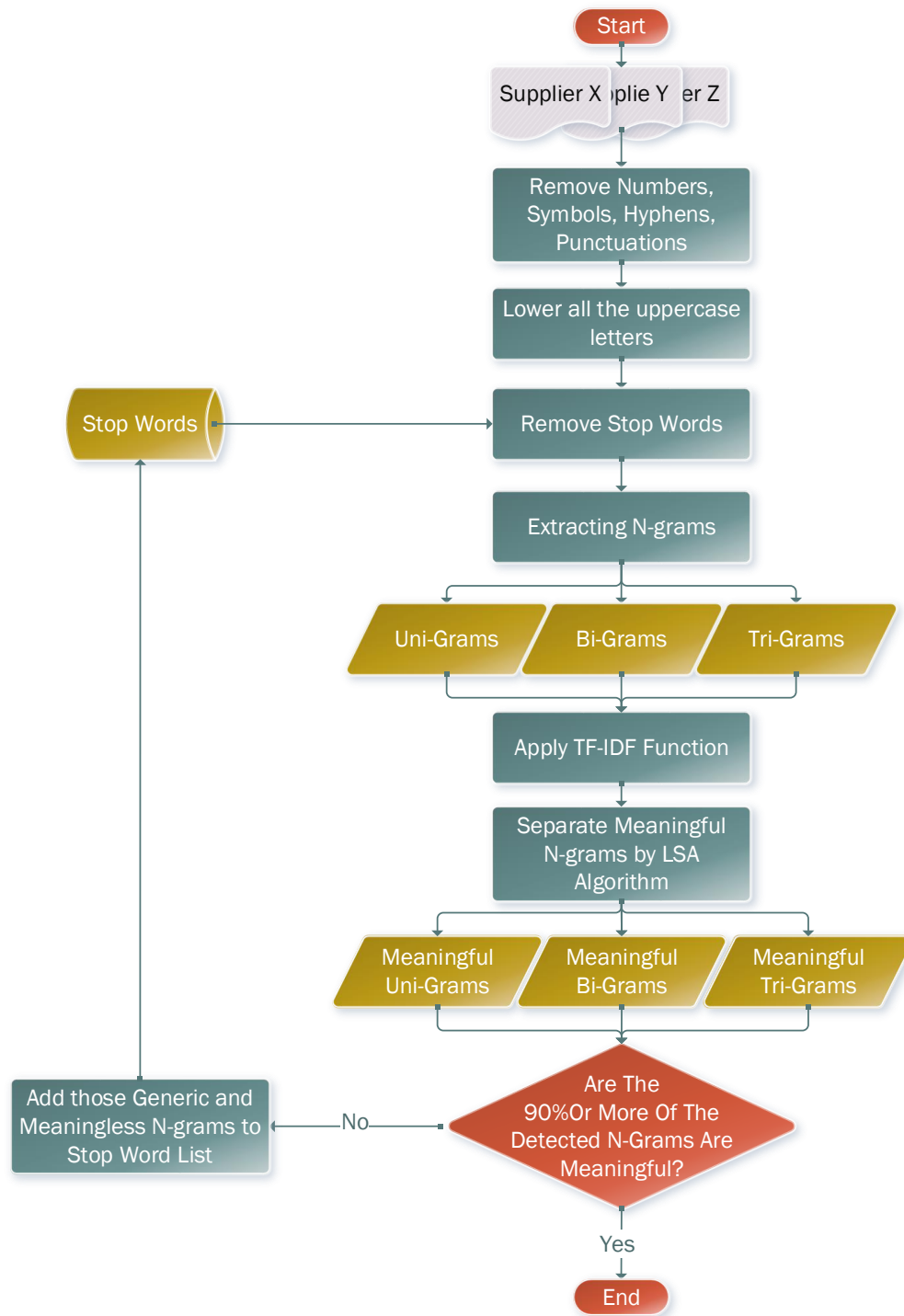
Some of the N-grams in the document-term matrix have higher distinguishing power than the others. Therefore, it is necessary to select only the N-grams that will be potentially influential when building supplier clusters. This also improves the computational efficiency of the algorithm when the size of matrix is too large. The following analyses can be conducted for this purpose:

- a) **N-gram's Normalized Frequency Analysis:** In this analysis, the summation of the normalized frequencies associated with each N-gram across all documents is

calculated and then the N-grams are sorted in a descending order. The top N-grams with highest normalized frequencies are selected. The cut-off line varies depending on where a significant gap appears in the sorted data. In the next step, the dataset with  $m$  suppliers is broken into  $m$  subsets. Each subset is a vector, which represents the N-grams of one distinct supplier and their associated frequencies. Then, inside each subset, the N-grams are sorted based on their frequencies so that we could detect the most important N-grams with highest frequencies for each supplier. This step is included to prevent the suppliers with lower terms and N-grams to be neglected in the total frequency method.

- b) Latent Semantic Analysis:** Latent Semantic Analysis (LSA) is a technique in natural language processing that can analyze the relationship between set of documents and terms by producing the set of concepts related to the documents and terms within documents (Landauer, 2006). LSA is typically used when the DTM is deemed too large or too noisy. The Singular Value Decomposition (SVD) is implemented for feature extraction and dimensionality reduction, which amounts to reducing the number rows and columns of the matrix (Golub & Reinsch, 1970). The SVD function allows us to use LSA to simultaneously increase the information density of each concept. The LSA method can detect the most significant N-grams based on their TF-IDF values (W. Zhang, Yoshida, & Tang, 2011).

Figure 13 shows the main steps regarding the process of cleaning the raw data, extracting the N-grams, and detecting the meaningful N-grams.



**Figure 13. Major Steps to Find Meaningful N-grams**

### 3.6. K-means Clustering

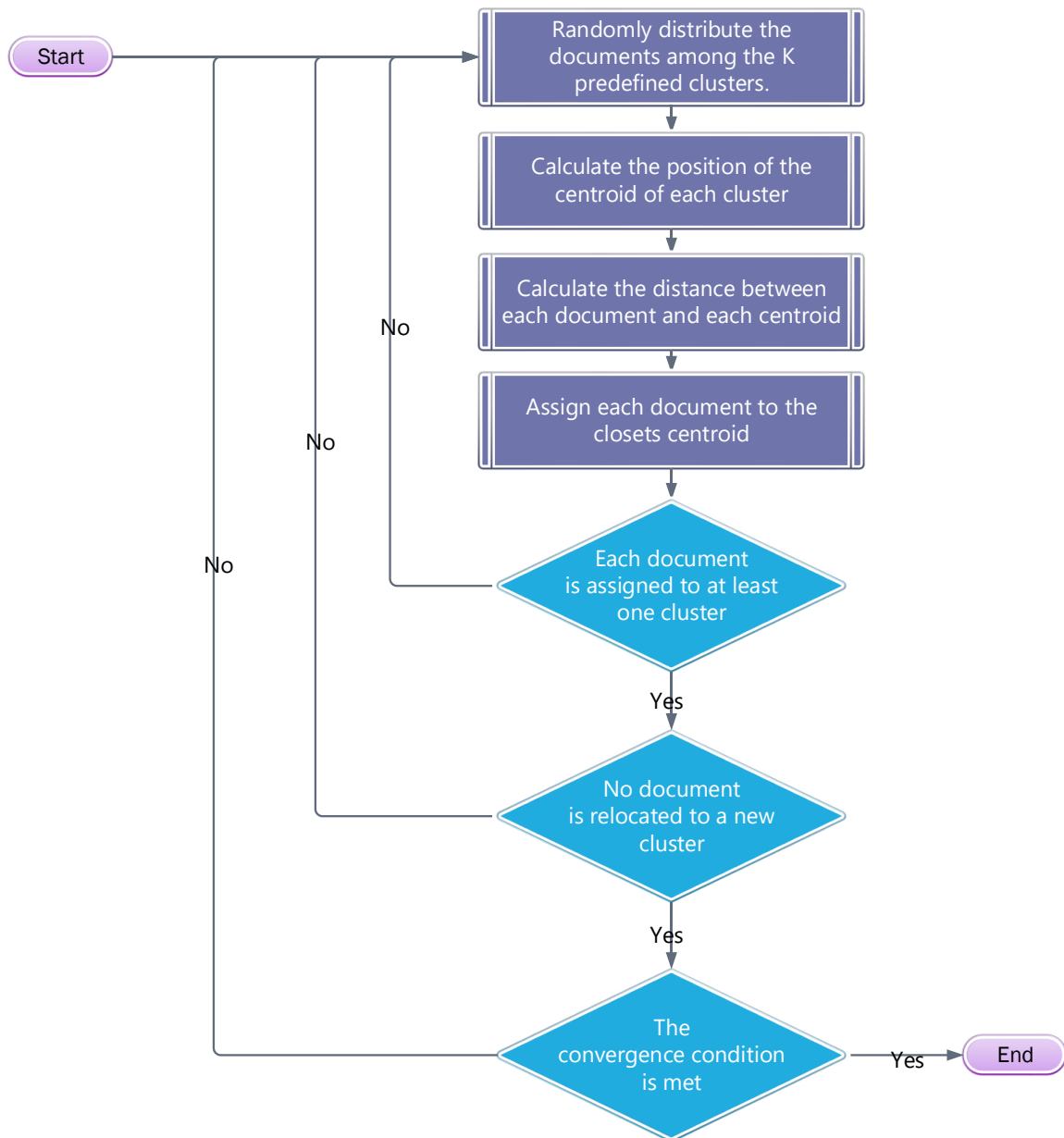
After all the preprocessing, adding N-grams to the dataset and lowering the rank of the N-gram matrix, the output csv file is used as the input to the clustering step. In this step, the goal is to create clusters or groups of similar documents (suppliers) inside the corpus. For this purpose, the *K-Means Clustering* algorithm is adopted which automatically divides the documents of the corpus into K groups such that documents in a cluster are more similar to each other in comparison with the documents in other clusters (Hartigan & Wong, 1979). In the next step, the algorithm defines one centroid per group. Each document is also assigned to the nearest centroid. The projection of multidimensional document term matrix on Euclidean planes, the distance of a specific document from the centroids of the clusters is calculated.

The goal of the k-means methodology is to minimize the sum of square of distances from documents (also known as data points) to the clusters' centroid. Finally, the algorithm goes through multiple iterations until the convergence condition is met. A set of observations ( $X_1, X_2, \dots, X_n$ ) each one representing a row (i.e. document) from the result of TF-IDF normalization step is assumed. Each observation contains “d” terms in it. In other words, it is a d-dimensional vector. The goal of K-Means clustering is to partition “N” observations into “K” ( $\leq N$ ) sets which is  $S = \{S_1, S_2, \dots, S_k\}$  in order to minimize the sum of squares within the clusters. Mathematically, the objective is to find:

$$\arg \min_S \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2 = \arg \min_S \sum_{i=1}^k |S_i| \text{Var } S_i \quad (3.3)$$

Where  $\mu_i$  is the mean of points in  $S_i$ .

Figure 14 shows the main steps of the K-means clustering algorithm. K is an integer number between one and the number of the documents. In order to have an efficient clustering method, the optimum value of K should be determined based on the size and diversity of the corpus (Kodinariya & Makwana, 2013).



**Figure 14. Main Steps of the Clustering Algorithm**

### 3.7. Topic Modeling

After the clusters are formed, the next step is to investigate the clusters in order to learn about the main theme and topic of the documents within each. However, the clustering method does not provide any description or characterization for the generated clusters. Therefore, the topic modeling technique is used in order to find the core topics for the clusters (Wallach, 2006).

Topic Modeling is an unsupervised text mining technique that can analyze the large amounts of unlabeled text. Latent Dirichlet Allocation (LDA) is used as the algorithm behind the topic modeling. The topic modeling is applicable on N-grams as well. LDA technique is used for automatically discovering topics in a group of unlabeled documents containing similar terms to that specific topic (Blei, Ng, & Jordan, 2003). A *topic* is a recurring pattern of words that frequently appear together. The number of topics should be determined before the experiments but the name and characteristics of the topics are unknown. In other words, each document is considered as a combination of one or more topics that already realized in the dataset.

The LDA model even shows that a word  $w$  in document  $d$  is assignable to one of the topics that have already been found in document  $d$ . Collapsed Gibbs sampling is one way the LDA learns the topics and the topic representations of each document (Yan, Guo, Lan, & Cheng, 2013). The basic steps of the LDA technique are listed below. First, inside each document, each word in the document should be assigned randomly to one of the  $K$  topics while  $K$  is already chosen. This random assignment results in topic representation for all documents in the dataset as well as word distributions for all the topics even

though they are not perfect. The following steps help the method to improve the accuracy from the initial assignment results.

1. For each document  $d$ , randomly allocate each word in the document to one of the  $t$  topics. This random allocation provides topic representations of all the documents as well as distributions of words of all the topics.
2. For each document  $d$ , calculate these two values:
  - a)  $P(\text{topic } t | \text{document } d)$ , which is the probability that specific word is assigned to topic  $t$  given that this word is inside the document  $d$ .
  - b)  $P(\text{word } w | \text{topic } t)$ , which is the probability of assignments to topic  $t$ , over all documents  $d$ , that come from word  $w$ .
3. Reassign the word  $w$  to a new topic  $t'$  where this new topic is chosen with the probability below in order to predict the probability that topic  $t'$  generated word  $w$ .
$$P(\text{topic } t' | \text{document } d) * P(\text{topic } t' | \text{document } d)$$
4. With repeating the last step, a steady state is achieved where the word-to-topic assignments becomes meaningful.

#### **4. Experiment**

In previous sections, the adopted unsupervised learning methodology was explained. In this section, the results of experimental validation of the proposed method is presented. The experiment has two parts. The first part of the experiment uses *seen* data to evaluate the accuracy of clustering process. In the second part of the experiment, K-means clustering and Topic Modeling techniques are applied to text extracted from unseen manufacturing suppliers. The objective of the second part of the experiment is to

find new clusters of manufacturing suppliers that are similar to each other with respect to manufacturing capabilities.

#### 4.1. Dataset Preparation

The dataset used in this experiment was collected from Thomas Net and mfg.com, two web-based sourcing portals, since they already contain multiple categories of custom manufacturing suppliers including heavy component machining and complex and difficult machining suppliers. For the first part of the experiment, 130 suppliers were selected from *heavy component machining* category and another 130 suppliers from *complex machining* category. Because the true category of each supplier is already known in this dataset, it is possible to evaluate the accuracy of the created clusters by identifying the *false positives*.

For the second part of this experiment, a data set composed of 150 unseen manufacturing suppliers that were randomly selected from Thomas Net, without any categorical preference, was used. For both seen and unseen datasets, the text from each supplier's website is imported into a single text document. All described preprocessing steps, TF-IDF function as well as dimensionality reduction steps are applied to the created documents. Figure 15 shows an excerpt associated with a supplier without any preprocessing.

```
> train$Text[21]
[1] "For 40 plus years, Coleys CNC has provided dozens of industries with superior quality machining ranging from simple production parts to complex machining of intricate designs, including 5th axis applications, and now high speed machining. We are dedicated to increasing capacity and growing our business by acquiring more of the most efficient machine tools and employing the best cutting strategies driven by the most high-tech software available. Through high speed machining we can reduce cycle time, sometimes as much as 40%. Thereby, ensuring the lowest possible prices for our customers without compromising quality. + years experience in high quality machining.\n\nCNC Turning Centers\nClick on photo for larger view.\nPage: 1 of 1\nVertical Turning Lathe\nBerthiez CNC Vertical Turning Lathe \nVTL with Live tooling. Turning to 50\" diameter, 50\" length.\nDaewoo Puma 400LB\nDaewoo Puma 400LB \nCNC Tu... <truncated>
```

**Figure 15. Plain Text from an Example Supplier in the Dataset**



## 4.2. Data Preprocessing

Once the dataset is created, the next step is to apply the necessary preprocessing steps such as removing the symbols, numbers, punctuations, and hyphens as well as lowering all the uppercase letters in the text. Finally, stop words are removed from the dataset.

"machining"	"center"	"toyoda"	"cnc"
"high"	"speed"	"bridge"	"machining"
"duty"	"cutting"	"large"	"parts"
"travels"	"toyoda"	"vertical"	"machining"
"vertical"	"machining"	"center"	"x"
"table"	"nikken"	"full"	"4th"
"speed"	"per"	"min"	"high"
"ways"	"cnc"	"programming"	"coley's"

**Figure 16. Example Supplier after Preprocessing**

Figure 16 shows the text associated with the example supplier after preprocessing.

## 4.3. Extracting N-grams from Supplier

For 260 suppliers, there are 135,312 distinct concepts that are only one-gram. If bigrams, trigrams, and more are added, more concepts could be detected in total. After adding the bigrams, the number of concepts are increased to 619,492 from 135,312 concepts. Bigrams, trigrams and generally N-grams could be added to the dataset and the TF-IDF transform the expanded feature matrix to see if accuracy (i.e. precision and recall) improves. Figure 17 shows some example of the tokenized data, which contain bigrams. Tokenization is the process of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens (Kaplan, 2005).

[421]	"run_tooling"	"tooling_fixtures"	"fixtures_jigs"
[424]	"jigs_assemblies"	"assemblies_prototypes"	"prototypes_cad"
[427]	"cad_cam"	"cam_services"	"services_include"
[430]	"include_plating"	"plating_heat"	"heat_treatment"
[433]	"treatment_anodizing"	"anodizing_coatings"	"coatings_edm"
[436]	"edm_welding"	"welding_capabilities"	"capabilities_include"
[439]	"include_research"	"research_development"	"development_reverse"
[442]	"reverse_engineering"	"engineering_working"	"working_models"
[445]	"models_industries"	"industries_served"	"served_include"
[448]	"include_alternative"	"alternative_energy"	"energy_manufacturers"
[451]	"manufacturers_medical"	"medical_defense"	"defense_use"
[454]	"use_fadal"	"fadal_vmc"	"vmc_3016fx"
[457]	"3016fx_oi"	"oi_control"	"control_can"
[460]	"can_simulate"	"simulate_program"	"program_prior"
[463]	"prior_machining"	"machining_programming"	"programming_can"

**Figure 17. Bigram Terms Detected from Example Supplier**

Figure 18 indicates a partial view of the trigrams extracted from the dataset.

[1]	"award_winning_custom"
[2]	"winning_custom_cnc"
[3]	"custom_cnc_precision"
[4]	"cnc_precision_machine"
[5]	"precision_machine_shop"
[6]	"machine_shop_founded"
[7]	"shop_founded_dave"
[8]	"founded_dave_cheryl"
[9]	"dave_cheryl_barrar"
[10]	"cheryl_barrar_don"
[11]	"barrar_don_dixon"
[12]	"don_dixon_work"
[13]	"dixon_work_closely"
[14]	"work_closely_partnership"
[15]	"closely_partnership_design"

**Figure 18. Trigram Terms Detected from Example Supplier**

#### 4.4. Normalization Based On TF-IDF

Figure 19 show the partial view of the document-term matrix after preprocessing steps. Each row represents a document (i.e. supplier) and each column represents a term detected from suppliers' websites. The numbers inside the matrix indicate the frequency of a term within a document.

	heavy	component	machining	services	include	turning	milling	materials	handled	powdered	metals	stainless	steel	tool	rene	high	nickel	steels	titanium	capabilities
text1	6	2	20	14	6	8	22	2	2	2	2	2	8	6	2	8	2	2	2	22
text2	9	2	42	18	6	17	22	11	0	0	10	10	25	6	0	4	0	4	0	34
text3	3	0	13	18	4	14	14	7	0	0	2	6	27	0	0	7	4	0	0	4
text4	3	5	23	8	2	4	13	3	0	0	0	2	37	4	0	7	0	0	3	12
text5	24	0	41	35	2	21	8	11	0	0	0	16	18	11	0	13	0	3	0	16
text6	1	0	19	2	1	2	4	1	0	0	0	3	6	5	0	12	0	0	0	2
text7	1	2	10	2	2	26	22	0	0	0	0	0	1	20	0	1	0	0	0	2
text8	0	2	5	7	0	1	2	2	0	2	0	0	0	0	0	0	0	0	0	2
text9	1	2	12	2	3	21	18	1	0	0	0	0	0	11	0	3	0	0	0	1

**Figure 19. Partial View of the Preprocessed Data before Applying TF-IDF for the Example Supplier**

Figure 20 shows the result of applying the TF-IDF function that normalizes the frequencies associated with the terms. According to Figure 19, the terms *machining* and *services* have the TF-IDF frequencies of zero. Since they appeared in all documents in the dataset, they are not important terms that can be used for distinguishing between several clusters.

	heavy	component	machining	services	include	turning	milling	materials	handled	powdered	metals	stainless	steel
text1	0.0006399771	0.0002960517	0	0	3.761249e-05	4.420069e-04	5.739252e-04	3.860692e-05	0.0009394626	0.00182675	0.0003686504	1.262238e-04	2.087001e-04
text2	0.0007972442	0.0002458687	0	0	3.123689e-05	7.800522e-04	4.766405e-04	1.763451e-04	0.0000000000	0.0000000000	0.0015308068	5.241397e-04	5.416370e-04
text3	0.0003854807	0.0000000000	0	0	3.020710e-05	9.318270e-04	4.399759e-04	1.627801e-04	0.0000000000	0.0000000000	0.0004441022	4.561743e-04	8.485249e-04
text4	0.0003501310	0.0008098481	0	0	1.371851e-05	2.418216e-04	3.710838e-04	6.336544e-05	0.0000000000	0.0000000000	0.0000000000	1.381139e-04	1.056162e-03
text5	0.0012028072	0.0000000000	0	0	5.890910e-06	5.451675e-04	9.806052e-05	9.976987e-05	0.0000000000	0.0000000000	0.0000000000	4.744636e-04	2.206362e-04
text6	0.0002461666	0.0000000000	0	0	1.446761e-05	2.550264e-04	2.408289e-04	4.455035e-05	0.0000000000	0.0000000000	0.0000000000	4.369670e-04	3.612433e-04
text7	0.0001156998	0.0003211346	0	0	1.359973e-05	1.558232e-03	6.225509e-04	0.000000e+00	0.0000000000	0.0000000000	0.0000000000	0.000000e+00	2.829777e-05
text8	0.0000000000	0.0009651945	0	0	0.000000e+00	1.801301e-04	1.701022e-04	1.258672e-04	0.0000000000	0.0000000000	0.0012018825	0.000000e+00	0.000000e+00
text9	0.0001901287	0.0005277181	0	0	3.352251e-05	2.068201e-03	8.370273e-04	3.440881e-05	0.0000000000	0.0000000000	0.0000000000	0.000000e+00	0.000000e+00
text10	0.0003112923	0.0004320088	0	0	3.659030e-05	6.449918e-04	3.045424e-04	0.000000e+00	0.0000000000	0.0000000000	0.0000000000	0.000000e+00	3.045424e-04

**Figure 20. A partial View of the Dataset after Applying TF-IDF Function for the Example Supplier**

#### 4.5. Detecting Meaningful N-grams from Heavy Machining Suppliers

The term-document matrix contains 619,492 N-grams at this stage. The sorting process consists of two separate steps. First, N-grams are sorted based on their total normalized frequencies across all suppliers. The N-grams in the upper 10% are selected as candidates for meaningful N-grams. In the second step, the dataset is broken into 130 subsets, each representing one distinct supplier as there are 130 heavy machining

suppliers in the corpus. In this step, the N-grams are sorted based on their frequencies within each supplier.

Similarly, top 10% of the N-grams of each subset are selected as candidates for meaningful N-grams. Finally, the candidates N-grams of both steps are reviewed by human expert to be identified as meaningful N-grams. Table 11 shows the top ten candidate N-grams resulted from total frequency analysis step. The candidate N-grams associated with each supplier could be identified and shown similar to the table 11.

**Table 11. Top 10 Candidate N-grams resulted from the Total Frequency Analysis**

N-GRAMS			
1	cnc machining center	6	turning
2	vertical milling	7	control tool changer
3	mazak	8	deep
4	heavy machining	9	large part
5	vertical	10	large shaft plant

The LSA method is used to reduce the dimensionality of the TF-IDF dataset. For this purpose, all the 130 heavy machining suppliers are selected and the LSA algorithm is applied. It has been mentioned before that the process of detecting the meaningful N-grams is an iterative process and the goal is to achieve the 90 percent purity of the meaningful N-grams. Table 12 shows the 25 most significant unigrams, bigrams, and trigrams representing the “Heavy Machining” class. At the end of this step, among more than 600,000 N-grams, less than 1,000 meaningful N-grams will be selected.

**Table 12. Top 25 Most Important N-Grams for Heavy Machining Suppliers Resulted by Latent Semantic Analysis Method**

#	Unigrams	Bigrams	Trigrams
1	flowforming	tool changer	contract manufacturing inc
2	stamping	shaft plant	chicago metal fabricators
3	discharge	metal fabricators	signal metal industries
4	cnc	ballast plow	metal industries inc
5	boring	transportation industry	shaft plant large
6	ballast	machine tool	heidenhain itnc control
7	filter	machine shop	control tool changer
8	turning	horizontal boring	contouring rotary table
9	mazak	engineering design	manual load unload
10	milling	industry manufacturing	cnc horizontal boring
11	screw	cnc machining	press brake forming
12	large	machined parts	elkins machine tool
13	trucking	cnc turning	cnc vertical lathe
14	crane	machining services	value added services
15	plow	laser cutting	turn cnc lathes
16	lift	metal fabrication	industry laser cutting
17	shaft	large milling	spindle x axis
18	fabrication	field services	brake forming laser
19	vertical	machining center	vertical machining center
20	acrylic	custom machining	milling machining services
21	hydraulics	milling services	deep hole boring
22	hole	repair maintenance	medium shaft plant
23	swing	precision machining	large shaft plant
24	forming	material handling	large milling plant
25	machine	heavy duty	width inside diameter

#### **4.6. K-Means Clustering Result on Suppliers of Heavy Machining and Complex Machining Classes**

For this dataset, K is set to 2 since the dataset was classified into two classes, namely, heavy machining and complex machining. It is assumed that all suppliers are correctly classified under heavy and complex machining classes by Thomas Net and mfg.com. As it can be seen in Table 13, most of the suppliers are classified correctly. The

overall precision is more than 98 percent. Since the accuracy condition is met, clustering is applied on unseen data in the next step.

**Table 13. K-means Results on Heavy Machining and Complex Machining Classes**

<b>True Positive (TP)</b> Number of heavy machining suppliers, which are classified as heavy machining class.	<b>129</b>	<b>0</b>	<b>False Negative (FN)</b> Number of heavy machining suppliers, which are not classified as heavy machining class.
<b>False Positive (FP)</b> Number of non-heavy machining suppliers, which are classified as heavy machining class.	<b>2</b>	<b>129</b>	<b>True Negative (TN)</b> Number of non-heavy machining suppliers, which are not classified as a heavy machining class.

#### **4.7. Clustering Result of Manufacturing Suppliers from Thomas Net with No Categorical Preference**

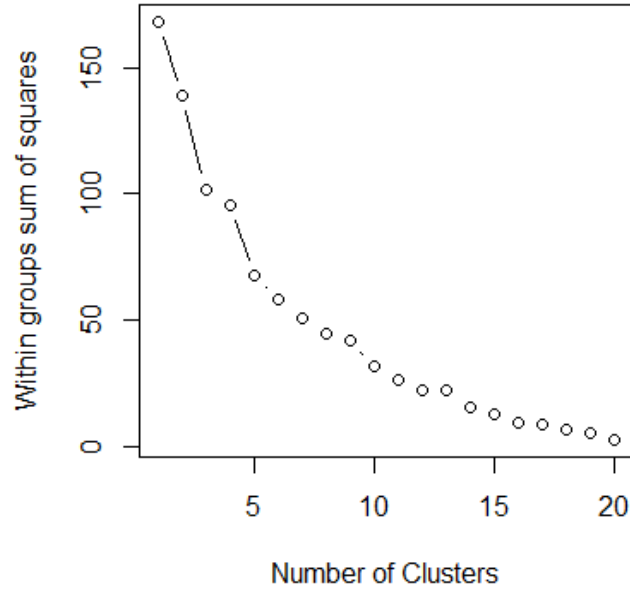
A dataset composed of 150 suppliers, randomly selected from Thomas Net, was formed for this step of the experiment. The K-means clustering technique is applied on these unseen suppliers in order to split the dataset into clusters. First, the optimum number of clusters should be determined. For this purpose, the sum of squares of distance between each document of a cluster and the centroid of the cluster is collected in each groups based on different number of clusters. Figure 21 shows this relationship. It is shown that the biggest reduction in sum of squares appears in first three to five clusters. Hence, the optimum number of clusters for this dataset is between three to five. Considering the fact that a relatively small dataset is used in this experiment, building three clusters seems more reasonable. Therefore,  $K$  is set to 3 as the number of clusters to be formed and the K-Means Clustering algorithm is run in order to group the unseen manufacturing suppliers into three clusters. 20 most important N-grams resulted by LSA method are shown in Table 14 to give an idea of the main features of the dataset in hand.

**Table 14. Top 20 Most Important N-Grams of Unseen Manufacturing Suppliers  
Dataset Resulted by Latent Semantic Analysis (LSA) Method**

N-GRAMS							
<b>1</b>	casting	<b>6</b>	machine	<b>11</b>	net shape forging	<b>16</b>	stamping
<b>2</b>	machining	<b>7</b>	mold	<b>12</b>	micro drilling	<b>17</b>	press forging
<b>3</b>	forging	<b>8</b>	die	<b>13</b>	electric discharge	<b>18</b>	screwswiss machining
<b>4</b>	milling	<b>9</b>	products	<b>14</b>	discharge milling	<b>19</b>	cnc machining
<b>5</b>	valves	<b>10</b>	steel	<b>15</b>	mold casting	<b>20</b>	casting

In K-Means clustering technique, the mean value of the normalized frequencies of each feature (N-gram) across all the documents within each cluster is calculated. In order to calculate the mean value, the frequencies are first converted to standard normal distribution to eliminate the difference in ranges of frequencies in the documents. The N-grams with the widest range (difference between min and max value) of mean value across all clusters are the most influential features when assigning a supplier to the cluster. On the other hand, if the mean value associated with a specific N-gram in different clusters are close to each other, then the N-gram does not play a significant role in assignment of suppliers to clusters.

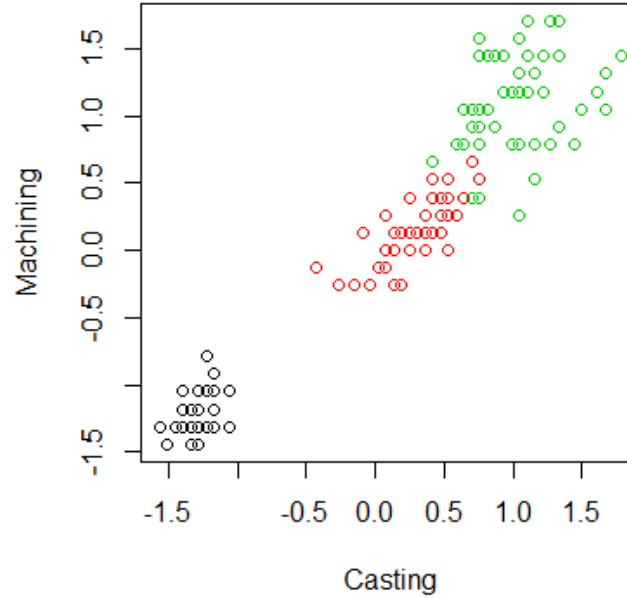
To identify the most distinguishing terms in cluster formation process, the N-grams and their mean values associated with normalized frequencies are listed for the three clusters. *Casting* and *Machining* terms are detected as the most distinguishing N-grams with highest difference in mean values between three clusters.



**Figure 21. Relationships between Number of Clusters and Within Groups Sum of Squares**

The clustering process resulted in formation of three clusters as shown in Figure 22. The vertical and horizontal axes in this figure indicate the normalized frequencies of the *casting* and *machining* terms in the dataset. After closely examining the websites associated with suppliers within each cluster, it was observed that the suppliers belong to three distinct categories, namely, machining, casting, and forging. The suppliers represented by black dots in Figure 22 provide precision machining services. There is no overlap between the machining cluster and the other two clusters. The casting and forging clusters have some partial overlap. To further investigate the topics of each cluster, Topic Modeling technique is used in the next step.

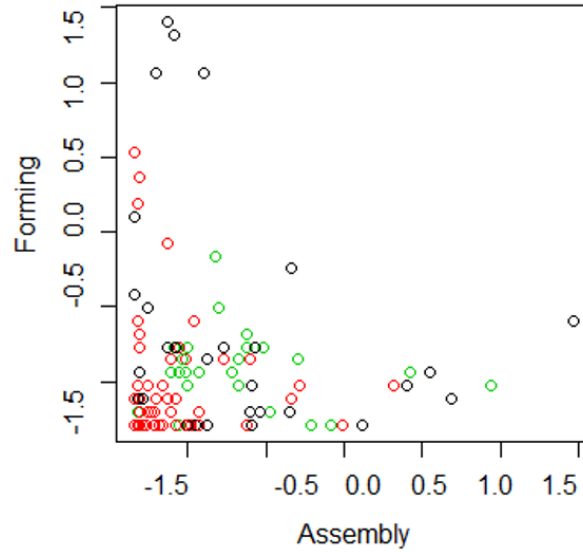




**Figure 22. The Result Map of the K-means Clustering Technique for 150 Suppliers Based on *Casting* and *Machining* Normalized Frequencies**

From the mean values of the frequencies of the N-grams across the suppliers, the mean value associated with *casting* and *machining* has the most difference, therefore, most distinguishing power compare to all other N-grams. Figure 22 also indicates that the *casting* and *machining* N-grams are appropriate N-grams to make the reasonable clusters.

On the other hand, if the mean values of the normalized frequencies associated with two N-grams are close to each other, it means that these two N-grams do not have a distinguishing power to create the meaningful clusters. For instance, mean value associated with *assembly* and *forming* N-grams are close to each other. Therefore, they could not make meaningful clusters in this specific dataset. Figure 23 shows the mapping of the K-means clustering technique based on *assembly* and *forming* normalized frequencies in each supplier. It is shown that the 3 clusters are not well-separated based on *forming* and *assembly* N-grams.



**Figure 23. The Result Map of the K-means Clustering Technique for 150 Suppliers  
Based on *Assembly* and *Forming* Normalized Frequencies**

#### **4.8. Topic Modeling Results**

Topic Modeling was applied to the entire dataset and the individual clusters in order to identify the representing features of the dataset and its clusters. As mentioned before, a *topic* is a group of co-occurring N-grams that best represent the information within the dataset. Table 15 shows the results of topic modeling when applied to the entire dataset with three identified topics. Only top 20 N-grams under each topic are shown in this table. From Table 15, one can infer that topic one represents the forging cluster, topic two is related to the casting and the third topic is associated with casting cluster. It should be mentioned that finding a one-to-one mapping between topics and clusters may not be possible for all datasets. The next step is to apply topic modeling the individual clusters.

**Table 15. Topic Modeling Results for 150 Unseen Manufacturing Suppliers**

#	Topic 1	Topic 2	Topic 3
1	press forging	screws	5 axis machining
2	coining	die casting	micro drilling
3	parts	molding	turning
4	quality	nuts	center
5	blanking	aluminum	manufacturing
6	shop	mold	assembly
7	stamping	washers	industries
8	upset	mold casting	electric discharge
9	manufacturing	anchors	swiss machining
10	turning	valves	5 axis milling
11	role forging	mold making	welding
12	services	steel	bearing
13	cnc machining	pump	vertical milling
14	industry	investment casting	milling turning
15	experience	structural	design
16	die forging	die making	reaming
17	heading	sand casting	cutting
18	die	centrifugal	custom manufacturing
19	fittings	foam	horizontal machining
20	drilling	bolts	edm

Table 16 shows the results of topic modeling when applied to the documents in the casting cluster. Again, the desirable number of topics was set to three. The N-grams under the generated topics collectively point to different processes, materials, and components related casting domain. This confirms the observation that suppliers in this cluster belong to the casting domain. Another utility of the created topics is providing some initial insights about associativity relationships between the terms in a topic. For example, the fact that the term *housing* (a type of component) is in the same cluster with

*sand casting, mold casting, and injection molding*, it can be inferred that these three processes are typically used to cast different types of housings based on the current state of casting industry. Finding semantic relationship between the extracted terms is a research problem that will be further investigated in the future.

**Table 16. Topic Modeling Result on the Cluster Related to Casting Suppliers**

#	Topic 1	Topic 2	Topic 3
1	centrifugal	mold making	nuts
2	structural	injection molding	iron
3	die making	molding	screws
4	pump	housing	bolts
5	casting services	aluminum	anchors
6	stamp die casting	mold	valves
7	plaster mold	sand casting	washers
8	lost foam casting	mold casting	steel
9	investment casting	anchors	milling
10	cast	vacuum	cutting

## 5. Conclusion

This article proposes an unsupervised learning method based on clustering and topic modeling for building groups of similar suppliers from the contents of their websites. This method can be applied to large volumes of unlabeled documents extracted from the Web. The unsupervised nature of the method eliminates the difficulties related to preparing training data. By building clusters of manufacturing suppliers and characterizing them through a set of features (terms and phrases), more structured manufacturing capability data will be generated. This can enhance the intelligence of sourcing and supply chain analysis solutions that operates on the content available on the Web.

A secondary outcome of this research is generating a vocabulary of manufacturing capability terms using Latent Semantic Analysis (LSA). The capability terms can be organized into more formal thesauri and knowledge graphs that can improve the performance of machine learning algorithms.

In the future, experimental validation will be conducted using a much larger dataset to obtain statistically significant results, more diverse clusters as well as more topics and their associated terms. This could result in discovering unknown patterns and novel trends in manufacturing domain. The feature selection method in the proposed framework is semi-automated in a sense that the meaningfulness for each N-gram needs to be determined manually by human expert in both methods, N-gram's Normalized Frequency Analysis and LSA. The method should detect the stop words automatically to eliminate the iterative process of detecting meaningless N-grams. Automated identification process of meaningful N-grams presents another research challenge to be addressed in the future.

## CHAPTER IV

### CONCLUSION AND FUTURE WORK

There is high volume of useful information hidden in suppliers' websites. In manufacturing area, suppliers provide valuable information about their capabilities such as the product they manufacture the processes they offer, and qualities they can achieve. Capability data published on suppliers' websites is often in unstructured format. Several approaches such as machine learning and data mining can be applied to organize the unstructured data and make it more usable. If manufacturing capabilities can be analyzed and evaluated based on the textual information provided on suppliers' websites, more informed decisions can be made when forming manufacturing supply chains.

To achieve this goal, there is a need for development of an automated text mining tool supported by analytical techniques. The *objective of this research* is to create a capability analysis framework for manufacturing supplier's classification through implementation of different supervised and unsupervised text mining techniques.

#### **1. Answers to Research Questions**

To provide the answers for the research questions identified in Chapter 1, different methods and algorithms were used throughout this study. In this chapter, the findings related to the research questions are summarized and the main contributions and the future works are discussed.

## **I. What is the most suitable text classification technique for capability-based supplier classification problem?**

Capability-based classification is a necessary first step for accurate supplier evaluation in decentralized scenarios. In Chapter II, a novel framework for capability-based supplier classification based on unstructured capability data is proposed. The proposed framework uses a concept-based method and is supported by a SKOS-based thesaurus referred to as Manufacturing Capability Thesaurus. The thesaurus encodes the domain knowledge and provides a semantically connected network of capability concepts. A formal thesaurus guides the feature selection process. Four classification techniques, namely, Nave Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM), were used. For all techniques, 70 percent of the data was randomly selected as training data and the rest was regarded as the test data. To eliminate the bias caused by the specific choice of training data, the classification was run for 10 times per technique for both scenarios. The results indicate that the SVM technique has the best performance among the used techniques with an average precision of 99%. As expected, Naive Bayes method was the low performer as was only used to provide a baseline for comparison. Another important observation was the overall precision improves when concept weightings are applied (scenario two). Also, it has shown that Bag of Concept method has higher accuracy (i.e. F-Measure) than Bag of Words method.

## **II. What are the steps needed for data preparation and cleaning?**

The plain text documents collected from manufacturing suppliers' website are converted into the csv format. Several preprocessing steps are applied to create a clean

dataset. The punctuations, hyphens, numbers, and symbols are removed and all the letters are converted to lowercase. Several stop words such as “can”, “also”, “yet”, “offer”, “capable”, “need”, “time”, “contact”, “include”, “available”, etc. are also applied to the both training and testing documents. The stemming function is avoided because in the manufacturing fields, different kinds of the words such as verb, noun, and adjectives, represent different concepts such as process, product, tool, and part and it is important to have the original words and keep their meanings.

Document-Term Frequency Matrix (DTM) is created based on the suppliers' corpus. Once the DTM is formed, the normalization functions are applied in order to improve the model efficiency. The use of Term Frequency-Inverse Document Frequency (TF-IDF) is a powerful technique for enhancing the information and signal contained within the document-frequency matrix. When comparing multiple documents, the longer documents would have higher individual term counts than the shorter documents. Since the term frequencies are so important in classification and clustering process, the TF function is used in order to eliminate the effect of document's length on data analysis process. In other words, TF is a function which normalizes the frequency of the terms based on the document's length. For this purpose, all the frequencies associated with the terms inside a specific document would be divided by the total number of frequencies of all terms in that document. It is obvious that the term which appears in all of the documents has no significant power to predict the class of the document. The IDF is function that normalizes the frequency of a term appearance in all documents. The multiplication of the TF and the IDF functions will consider both frequencies with regards to the document's length as well as the frequencies of the concepts in all other



documents is the corpus. In large documents, in two ways the result could be improved. The frequencies could be normalized based of the total word count in the document. The TF function is used for this reason. Also, the words that appear in all documents should be penalized and the IDF is used for this function.

### **III. How to manually create the corpus and manipulate it as training data in data mining techniques?**

The raw text which is available in each manufacturing supplier's website is copied and pasted to a separate text file. Each text file represents a manufacturing supplier's text. In order to have a more relevant data, the "service" page, the "capability" page or other pages related to the supplier's capability is captured. The plain text documents are converted into the csv format. In order to select the most appropriate training data for the classification of manufacturing suppliers, a function named "CreateDataPartition" inside of the *caret* package (short for classification and regression training) in R is adopted to create a stratified data partition which assign 70% of the data to training data and 30% of the data to test data. The *caret* package contains functions to streamline the model training process for complex regression and classification problems (Kuhn & others, 2008). Once the data is split, the preprocessing is started for both training and test data.

### **IV. How text classification techniques can help organize suppliers based on their capabilities?**

In text classification problems, one of the key steps is selecting the features that can uniquely characterize a class of interest. *Feature selection* can be done either

manually or automatically through machine learning. Manual feature selection entails manual selection of the features for each class by domain experts. Manual feature engineering often requires considerable effort, and the selected features might be subjective and problem-specific. Automated feature engineering requires less time and effort compared to the manual method but its main drawback is that it often ignores the semantic dependencies between the features. In Chapter II, a semi-automated technique, guided by a formal thesaurus, is adopted for feature engineering. In the presence of a thesaurus, one can dynamically define various capability classes simply by selecting the relevant terms, or features, available in the thesaurus for the class of interest. The feature selection process is guided by the underlying semantic model of the thesaurus. This eliminates the time-consuming and costly steps required for creating gold standard training corpus for each class of interest that is often used in supervised classification techniques based on machine learning. The formal nature of the proposed thesaurus enhances the semantic relevance of the results. The proposed classification methodology is validated experimentally through forming two capability classes, namely, heavy component machining and difficult and complex machining. It can be inferred from the classification methodology that manufacturing suppliers could be classified under diverse classes based on the requirements and demands.

## **V. What types of hidden knowledge patterns exist in manufacturing suppliers' websites?**

This question is answered in Chapter III. A dataset composed of 150 suppliers, randomly selected from Thomas Net. The K-means clustering technique is applied on these unseen suppliers in order to split the dataset into clusters. First, the optimum

number of clusters is determined. For this purpose, the sum of squares of distance between each document of a cluster and the centroid of the cluster is collected in each groups based on different number of clusters. It has shown that the optimum number of clusters for the dataset is three. Therefore,  $K$  is set to 3 as the number of clusters to be formed and the K-Means Clustering algorithm is run in order to group the unseen manufacturing suppliers into three clusters. The most important N-grams resulted by LSA method have shown to give an idea of the main features of the dataset in hand. *Casting* and *Machining* terms are detected as the most distinguishing N-grams with highest difference in mean values between three clusters. After closely examining the websites associated with suppliers within each cluster, it was observed that the suppliers belong to three distinct categories, namely, machining, casting, and forging. Topic Modeling was applied to the entire dataset and the individual clusters in order to identify the representing features of the dataset and its clusters. It can be inferred that the topics represent the forging, casting and the casting clusters. It should be mentioned that finding a one-to-one mapping between topics and clusters may not be possible for all datasets.

Topic modeling is also applied to the documents in the casting cluster. Again, the desirable number of topics was set to three. The terms under the generated topics collectively point to different processes, materials, and components related casting domain. This confirms the observation that suppliers in one of the clusters belong to the casting domain. Another utility of the created topics is providing some initial insights about associativity relationships between the terms in a topic. For example, the fact that the term *housing* (a type of component) is in the same cluster with *sand casting*, *mold casting*, and *injection molding*, it can be inferred that these three processes are typically

used to cast different types of housings based on the current state of casting industry.

Finding semantic relationship between the extracted terms is a research problem that will be further investigated in the future.

## **VI. What are the important terms which suppliers, in contract manufacturing industry, use in order to describe their capabilities?**

In order to find the most frequent terms which is used in manufacturing suppliers' website, two methods are introduced in Chapter III. Some of the N-grams in the document-term matrix have higher distinguishing power than the others. Therefore, it is necessary to select only the N-grams that will be potentially influential when building supplier clusters. This also improves the computational efficiency of the algorithm when the size of matrix is too large. The following analyses can be conducted for this purpose.

**N-gram's Normalized Frequency Analysis:** In this analysis, the summation of the normalized frequencies associated with each N-gram across all documents is calculated and then the N-grams are sorted in a descending order. The top N-grams with highest normalized frequencies are selected. The cut-off line varies depending on where a significant gap appears in the sorted data. In the next step, the dataset with  $m$  suppliers is broken into  $m$  subsets. Each subset is a vector, which represents the N-grams of one distinct supplier and their associated frequencies. Then, inside each subset, the N-grams are sorted based on their frequencies so that we could detect the most important N-grams with highest frequencies for each supplier. This step is included to prevent the suppliers with lower terms and N-grams to be neglected in the total frequency method.

**Latent Semantic Analysis:** Latent Semantic Analysis (LSA) is a technique in natural language processing that can analyze the relationship between set of documents and terms by producing the set of concepts related to the documents and terms within documents. LSA is typically used when the DTM is deemed too large or too noisy. The Singular Value Decomposition (SVD) is implemented for feature extraction and dimensionality reduction, which amounts to reducing the number rows and columns of the matrix. The SVD function allows us to use LSA to simultaneously increase the information density of each concept. The LSA method can detect the most significant N-grams based on their TF-IDF values.

Both methods are applied in order to find the most distinguishing as well as most frequent terms in the manufacturing suppliers' websites.

## **2. Contribution**

The contributions of this thesis can be discussed from methodology as well as tool and information model development perspectives.

### **2.1. Methodological Contributions**

In the first part of this research (i.e. supervised learning), a novel framework for capability-based supplier classification based on unstructured capability data is proposed. The proposed framework uses a concept-based method and is supported by a SKOS-based thesaurus referred to as Manufacturing Capability Thesaurus (MC Thesaurus). This is a new framework for document classification that can be applied to different classification scenarios. Also, this work is one of the first demonstrations of how a

semantic thesaurus can enhance the precision of machine learning techniques such as document classification.

## **2.2. Information Models and Tools**

Development of the thesaurus, by itself, is a notable contribution since it is the first formal thesaurus of its kind in the manufacturing domain. The MC Thesaurus encodes the domain knowledge and provides a semantically connected network of capability concepts and guides the feature selection process. The MC Thesaurus is developed in a bottom-up fashion through tagging the key terms on suppliers' websites. The novel feature of the MC Thesaurus is that it contains the informal terms typically used in contract manufacturing industry. The MC Thesaurus can be extended and validated by domain experts in a collaborative fashion using cloud-based tools. Therefore, it can serve as a trustable source of manufacturing capability knowledge.

Furthermore, since it is based on SKOS, it can be linked to various open-source datasets on Linked Open Data (LOD) and reuse the existing concept models, thus enabling continuous and dynamic evolution and extension. Additionally, because SKOS thesauri are machine-understandable, the MC thesaurus can be integrated seamlessly with different semantic solutions that can support supply chain decisions. The crowdsourced extension and validation of MC Thesaurus significantly reduces the cost of evolution and validation. The proposed document classification framework is generic enough that can be applied to other areas such as customer review analysis or maintenance report classification.

Another contribution is development of the SKOS Tool. The SKOSTool is a user-friendly website programmed by JAVA in INFONEER lab<sup>5</sup> in order to facilitate the process of text analytics in this research and increase the efficiency of the supervised learning experiment. Some of the important capabilities associated with the SKOSTool are analyzing the corpus of text data, creating concept models for target classes, and detecting the frequencies of concepts in the websites.

The second part of this research proposes an unsupervised learning method based on clustering and topic modeling for building groups of similar suppliers from the contents of their websites. This method can be applied to large volumes of unlabeled documents extracted from the Web. The unsupervised nature of the method eliminates the difficulties related to preparing training data. By building clusters of manufacturing suppliers and characterizing them through a set of features (terms and phrases), more structured manufacturing capability data will be generated. This can enhance the intelligence of sourcing and supply chain analysis solutions that operates on the content available on the Web. A secondary outcome of this research is generating a vocabulary of manufacturing capability terms using Latent Semantic Analysis (LSA). The capability terms can be organized into more formal thesauri and knowledge graphs that can improve the performance of machine learning algorithms.

### **3. Future Work**

In the future, experimental validation will be conducted using a much larger dataset to obtain statistically significant results, more diverse clusters as well as more

---

<sup>5</sup> <http://infoneer.wp.txstate.edu/>

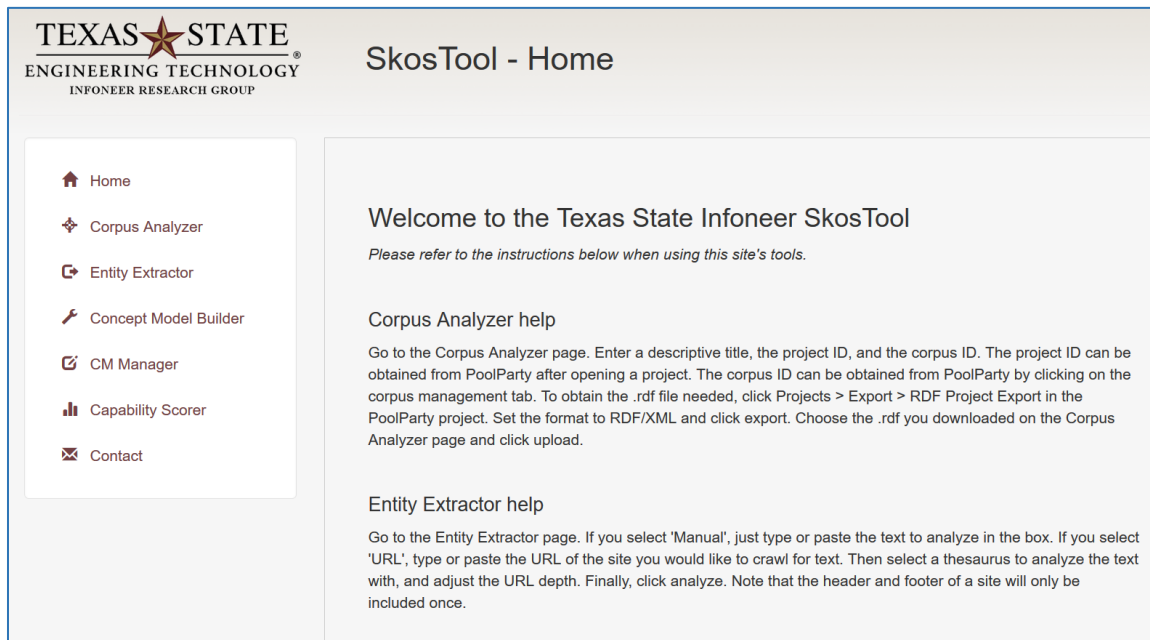
topics and their associated terms. This could result in discovering unknown patterns and novel trends in manufacturing domain. The feature selection method in the proposed framework is semi-automated in a sense that the meaningfulness for each N-gram needs to be determined manually by human expert in both methods, N-gram's Normalized Frequency Analysis and LSA. The method should detect the stop words automatically to eliminate the iterative process of detecting meaningless N-grams. Automated identification process of meaningful N-grams presents another research challenge to be addressed in the future. Finally, capability modeling and quantification needs to be done through capability scoring and supplier scoring and ranking based on their capabilities.



## APPENDIX A

### SKOSTool Website

In this research, Simple Knowledge Organization System (SKOS) is used for representation of manufacturing capability vocabulary. SKOS has multiple advantages. It is W3C Standard and light-weight semantic model. It also provides the options for expansion of relationship types, grouping structures, and mapping properties. In this section, the description associated with the SKOSTool website, which is implemented through this research by JAVA language, is provided. The website is linked with the Manufacturing Capability Thesaurus. The general and detailed descriptions for multiple sections will be presented. Figure 24 shows the Screenshot from the home page of SKOSTool Website.



**Figure 24. Home Page of SKOSTool Website**

Followings are services that The SKOSTool website offers:

- Corpus Analyzer
- Entity Extractor
- Concept Model Builder
- Concept Model Manager
- Capability Scorer

### **1. Concept Model Builder**

This section is to create the concept model for target classes in supervised learning method. In this section the steps will be explained for *Complex and Difficult Machining* class.

First off, the entry concepts associated with the *Complex Machining* class should be determined and entered by user. The entry concepts could be *complex machining*, *difficult machining*, *live tooling*, etc. After determining the entry concepts, multiple SPARQL queries are run to find the broader, narrower, and related concepts as well as alternative labels of those concepts. The user can choose to include the related concepts, alternative labels or top-level concepts or not. Also, the user may specify the narrowing and broadening level of the queries.

Figure 25 shows a screenshot from this step. For instance, *live tooling* is added as an entry concept for *Complex Machining* class.

Select concept #1:

Live Tooling

☒ Include related concepts
 ☒ Include alternative labels
 ☐ Include top-level concepts

Narrowing levels:  2  
 Broadening levels:  1

+ Add concept

Submit

**Figure 25. Screenshot from Selection of the Entry Concepts**

Once all the entry concepts as selected and the options are specified, several queries are run and all of the important concepts associated with specific class is collected. Finally, once the appropriate weightings are assigned into the important concepts through the weighting method, the concept model is ready. Figure 26 shows partial view of the concept model for *complex machining* class.

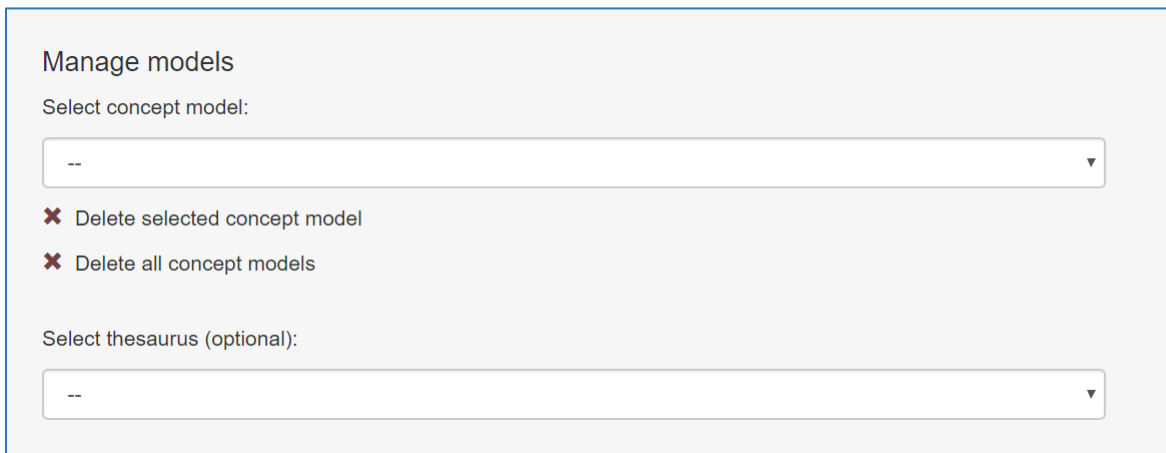
Concept:	Weight
5 axis machining	1
5 sided milling	1
5-Axis Machining	5
5-Axis machining capability	1
5-Axis Simultaneous Machining	1
7-Axis Machining	5
Complex CNC Machining	5
complex component	1
complex dimensional shapes	5
complex drilling problems	5
Complex Geometries	1
complex machined components	1
complex machined parts	5
Complex Machining	9
complex metal shape	1

**Figure 26. Partial View of the Important Concepts of Complex Machining Class**

Each target class has its own associated concept model. In the next section, the SKOSTool's ability in order to modify the concept model will be introduced.

## 2. Concept Model Manager

The Concept Model Manager section can be used in order to edit the concept model resulted by Concept Model Builder section. In this section of the SKOSTool, the irrelevant concepts in case of incorrect query results could be removed or missed concepts in concept model of specific class could be added to the concept model. Figure 27 show the ability of CM Manager in deleting the concepts from concept model.



Manage models

Select concept model:

--

✖ Delete selected concept model

✖ Delete all concept models

Select thesaurus (optional):

--

**Figure 27. Ability to Delete One or More Concepts from Concept Model**

Figure 28 shows the ability of modifying the concepts in concept model manager.

Add labels from thesaurus to model (default weight of 1):

Delete labels from model:

**Figure 28. Modifying the Concepts inside the Concept Model**

Concept Model Manager also has the ability to adjust the weighting associated with concepts in the concepts model. Figure 29 show the partial view of the concept model of the *complex machining* class which could be adjusted manually.

Edit weights:

<input type="text" value="1"/>	5 axis machining
<input type="text" value="1"/>	5 sided milling
<input type="text" value="5"/>	5-Axis Machining
<input type="text" value="1"/>	5-Axis machining capability
<input type="text" value="1"/>	5-Axis Simultaneous Machining
<input type="text" value="5"/>	7-Axis Machining
<input type="text" value="5"/>	Complex CNC Machining
<input type="text" value="1"/>	complex component
<input type="text" value="5"/>	complex dimensional shapes
<input type="text" value="5"/>	complex drilling problems
<input type="text" value="1"/>	Complex Geometries
<input type="text" value="1"/>	complex machined components
<input type="text" value="5"/>	complex machined parts
<input type="text" value="9"/>	Complex Machining
<input type="text" value="1"/>	complex metal shape

**Figure 29. Adjusting the Weightings inside the Concept Model**

### 3. Entity Extractor

This section is used to collect the text from suppliers and prepare the test data for the text analytics. Text could be entered in two ways, manually and by website's URL. When dealing with URL, the number of levels of depth for web crawling should be specified beforehand. The user could also determine to include zero-occurrence concepts and top-level concepts or not. Figure 30 shows the screenshot for the text input steps in entity extractor.

The screenshot shows a web interface titled "Upload text". It features two radio buttons: "Manual" (unselected) and "URL (case sensitive)" (selected). Below the radio buttons is a large text input field. Underneath the input field is a label "Select thesaurus:" followed by a dropdown menu showing "MCT". Below the dropdown are four options with checkboxes: "Upload new thesaurus" (with an upload icon), "Delete all thesauruses" (with a delete icon), "Include zero-occurrence concepts" (unchecked), and "Include top-level concepts" (unchecked). Below these is a checked checkbox for "Show URL preview page". At the bottom, there is a "URL depth:" label followed by a slider control set to "1". A grey "Analyze" button is located at the bottom left of the form.

**Figure 30. Input Text**

The URL preview check box is also provided for users to see all of the resulted URL and make sure that all of the links are useful and relevant. Irrelevant URL links could be unchecked and removed. Figure 31 shows the URL preview step in entity extractor.

### Preview

*32 links have been gathered. Please uncheck the links you don't want the crawler to parse and then click submit.*

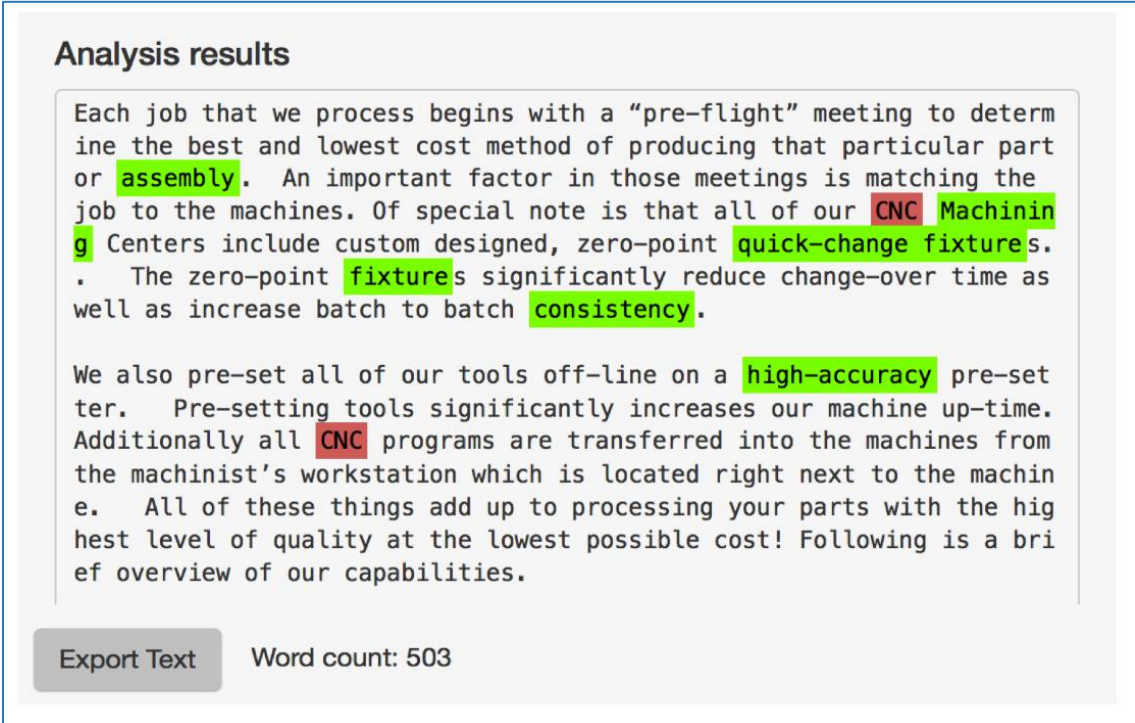
- ☒ <http://cncind.com>
- ☒ <http://cncind.com/aerospace-machining-2/>
- ☒ <http://cncind.com/aerospace-machining/>
- ☒ <http://cncind.com/announcing-new-director-engineering/>
- ☒ <http://cncind.com/announcing-new-production-manager/>
- ☐ <http://cncind.com/blog/>
- ☒ <http://cncind.com/capabilities-2/>
- ☒ <http://cncind.com/cnc-industries-as9100-certification/>
- ☒ <http://cncind.com/cnc-industries-awarded-new-multi-year-contract/>
- ☐ <http://cncind.com/contact/>
- ☒ <http://cncind.com/custom-machining/>
- ☒ <http://cncind.com/employment-opportunities/>
- ☒ <http://cncind.com/equipment-list/>
- ☐ <http://cncind.com/faq/>
- ☒ <http://cncind.com/green-manufacturing/>
- ☒ <http://cncind.com/guiding-principles-for-success/>
- ☒ <http://cncind.com/haas-vf-5-press-release/>
- ☒ <http://cncind.com/materials-machined/>
- ☒ <http://cncind.com/naics/>

**Figure 31. URL Preview**



For instance in this example, three links related to *blog*, *contact*, and *frequently asked questions* sections are unchecked and removed from the final text file.

Once the relevant links are selected and web crawling is finished, the text associated with all of the links inside each supplier is merged and saved as a single text file. Preferred label of the important concepts are highlighted in green and alternative labels are highlighted in red. Figure 32 shows the screenshot for analysis result in entity extractor section. The text file could be exported as a .txt format.



**Analysis results**

Each job that we process begins with a “pre-flight” meeting to determine the best and lowest cost method of producing that particular part or **assembly**. An important factor in those meetings is matching the job to the machines. Of special note is that all of our **CNC Machining** Centers include custom designed, zero-point **quick-change fixtures**. The zero-point **fixtures** significantly reduce change-over time as well as increase batch to batch **consistency**.

We also pre-set all of our tools off-line on a **high-accuracy** pre-setter. Pre-setting tools significantly increases our machine up-time. Additionally all **CNC** programs are transferred into the machines from the machinist’s workstation which is located right next to the machine. All of these things add up to processing your parts with the highest level of quality at the lowest possible cost! Following is a brief overview of our capabilities.

**Export Text**      Word count: 503

**Figure 32. Analysis Result**

A vector model containing the important concepts of the target class as well as their associated frequencies inside the supplier. The vector model could be sorted alphabetically or could be sorted by concepts’ frequencies. Figure 33 shows the partial

view of the vector model from an example supplier. The output of the entity extractor is this vector model. The vector model could be exported and saved by CSV format.

Sort table:

☐ Alphabetical ☒ Occurrences

Concept (preferred label)	Occurrences
CNC Machining Center	15
Machining	7
vertical machining center	4
Computer-aided Manufacturing	4
Fixture	3
SolidWorks	2
quick-change fixture	2
STEP	1
Quality Assurance	1
high-speed spindle	1
high-accuracy	1

**Figure 33. Vector Model from an Example supplier**

The capability scorer will be used in future work.

## APPENDIX B

### R Codes and Instructions for Supervised Learning Method

In this section, the instructions and R codes is provided in order to conduct the supervised learning experiment. The instruction consist of collecting and reading the data, applying some preprocessing functions, and finally run four classifications techniques. In the supervised learning experiment, the data is in CSV format and most of the preprocessing steps are applied in the SKOSTool website and the output of the concept model builder and entity extractor sections are clean data. But, for the bag of words approach, the experiment is started with raw text in XML format.

Let's start with the bag of words approach. In this experiment, several packages such as *tm*, *e1071*, *xml*, and *wordcloud* are used. Here is an example of the text of the supplier in XML format.

```
<?xml version="1.0" encoding="UTF-8"?>
<Info>
<Type>Heavy</Type>
<text>
Heavy and heavy component machining services include turning and milling services. Materials handled include powdered metals, stainless steel, tool steel, Rene®, high nickel steels and titanium. Capabilities also include drilling, deep hole boring, manual planning, welding and rolling. ISO 9001:2008 certified woman-owned custom manufacturer service company offering CNC machining, CNC milling, CNC turning, machined castings 4th 5th axis rotary transfers. Secondary operations include stamping, welding, centerless grinding, double disc grinding, sandblasting vibratory finishing.
</text>
</Info>
```

After reading the text file, several preprocessing steps such as removing numbers, punctuations and symbols are applied and stemming function is avoided. Multiple stop words are added to the stop word list to improve the purity and relevance of the text. After the preprocessing, the text files are partitioned into training and test data. Usually 70 to 80 percent of the data is selected as training data. Finally, the Naïve Bayes classifier is used to classify the suppliers into heavy and complex machining classes. Figure 34-35 shows the R code associated with the Bag of Words (BOW) approach along with the appropriate comments.

```

1 library(tm)
2 library(e1071)
3 library(XML)
4 library(wordcloud)
5 setwd("C:/Corpus")
6
7 #define function to read XML, specify content as 'content' in the latest {tm} version, but as 'Content' in older versions
8 readCorpus <- readXML(spec=list(Type=list("node","/Info/Type"), content=list("node","/Info/text")),doc=PlainTextDocument())
9
10 #load the corpus from the manufacturing files, specify UTF-8 encoding
11 Corpus.Heavy <- Corpus(DirSource("C:/Corpus/Heavy", encoding="UTF-8"), readerControl=list(reader=readCorpus))
12 Corpus.Complex <- Corpus(DirSource("C:/Corpus/Complex", encoding="UTF-8"), readerControl=list(reader=readCorpus))
13
14 Corpus <- c(Corpus.Heavy, Corpus.Complex)
15
16 for(i in 1:length(Corpus))
17 { Corpus[[i]]$content <- tolower(Corpus[[i]]$content) }
18
19 compound<- c("arbor support", "automatic bar feeder", "automatic chucking lathe","cnc Heavy", "automatic lathe ",
20 "bar feeder", "bar feeders", "cnc lathe ", "cnc Complex", "contour Complex", "cut off", "diamond Complex",
21 "drilling and threading", "dual spindle cnc kathe", "end facing", "engine lathe", "face grooving",
22 "fay automatic lathe", "form Complex ", "geometric lathe", "hard Complex", "horizontal boring",
23 "horizontal Complex", "hot water drill", "internal threading", "lathe work", "manual machining",
24 "manual Complex", "off axis Complex", "off center Complex", "ornamental Complex", "rose engine lathe",
25 "screw machine ", "screw machining", "single point cutting", "single point diamond Complex",
26 "straight Complex", "swiss precision machining", "swiss Complex", "swiss type lathe",
27 "swiss type Complex", "taper Complex", "Complex center", "turret lathe", "vertical boring",
28 "vertical Complex", "7 axis contour Heavy", "arbor Heavy", "climb Heavy", "cnc horizontal Heavy",
29 "cnc vertical Heavy", "contour Heavy ", "conventional Heavy", "down Heavy", "drilling and threading",
30 "duplex Heavy", "end Heavy", "face grooving", "face Heavy", "fly cutting", "gang Heavy",
31 "gang Heavy operation", "gear cutting", "gear hobbing", "hard Heavy", "high speed Heavy",
32 "horizontal Heavy", "hot water drill", "manual machining", "manual mill", "manual Heavy",
33 "manual millis", "multi point cutting", "pencil Heavy", "shell Heavy", "side Heavy", "slab Heavy",
34 "spot face", "spot facing", "straddle Heavy", "thread Heavy", "thread whirling", "up Heavy", "vertical Heavy")
35
36 for (j in seq(Corpus))
37 {
38   for (i in seq(compound)){
39     Corpus[[j]]$content <- gsub(compound[i], gsub(pattern = " ", replacement = "", compound[i]), Corpus[[j]]$content)
40   }
41 }

```

**Figure 34. R Codes for Bag of Word Approach -Naive Bayes Technique**

```

43 #prepare the documents for classification
44 #toString <- content_transformer(function(x, from, to) gsub(from, to, x))
45 Corpus <- tm_map(Corpus, stripWhitespace)
46 Corpus <- tm_map(Corpus, removePunctuation)
47 Corpus <- tm_map(Corpus, removeWords, stopwords("english"))
48 Corpus <- tm_map(Corpus, removeNumbers)
49 Corpus <- tm_map(Corpus, stripWhitespace)
50 Corpus <- tm_map(Corpus, removeWords, c("steel", "the", "wire", "available", "services", "hardware", "is", "product", "capable",
51 "markets", "industry", "iso", "lbs", "delivery", "provide", "design", "commercial",
52 "services", "materials", "available", "various", "quality", "parts"))
53
54 Corpus <- tm_map(Corpus, removeWords, "steel", "Custom", "wire", "available", "service", "cnc")
55 my_stopwords <- c(stopwords("english"), "services", "and", "custom", "also", "materials", "heat", "available", "including", "iso", "lbs",
56 "wire", "hardware", "markets", "capable", "product", "design", "provide", "delivery", "material",
57 "industry", "commercial", "part", "parts", "available", "quality", "include", "capabilities")
58
59 Corpus <- tm_map(Corpus, removeWords, my_stopwords)
60 Corpus <- tm_map(Corpus, stemDocument) #new
61
62 # for an in-depth explanation of this see the Naive Bayes methods
63 #Corpus.train <- c(sample(1:5), sample(21:25))
64 Corpus.train <- c(20, 21, 1, 2, 24, 42, 23)
65
66 Corpus.train.dtm <- DocumentTermMatrix(Corpus[Corpus.train])
67 Corpus.train.dtm <- removeSparseTerms(Corpus.train.dtm, 0.85)
68 Corpus.dict <- dimnames(Corpus.train.dtm)[[2]]
69 Corpus.train.dtm.bin <- inspect[Corpus.train.dtm]
70 Corpus.train.dtm.bin <- Corpus.train.dtm.bin > 0
71 Corpus.train.dtm.bin <- as.data.frame(Corpus.train.dtm.bin)
72 for(i in 1:length(Corpus.train.dtm.bin)){ Corpus.train.dtm.bin[,i] <- as.factor(Corpus.train.dtm.bin[,i])}
73
74 #prepare test data
75 Corpus.test.dtm <- DocumentTermMatrix(Corpus[-Corpus.train], list(dictionary=Corpus.dict))
76
77 Corpus.test.dtm.bin <- inspect(Corpus.test.dtm)
78 Corpus.test.dtm.bin <- Corpus.test.dtm.bin > 0
79 Corpus.test.dtm.bin <- as.data.frame(Corpus.test.dtm.bin)
80 for(i in 1:length(Corpus.test.dtm.bin)){ Corpus.test.dtm.bin[,i] <- as.factor(Corpus.test.dtm.bin[,i]) }
81
82 #prepare class labels
83 Corpus.lab <- as.vector(unlist(lapply(Corpus, meta, tag="Type")))
84 Corpus.lab <- as.factor(Corpus.lab)
85
86 #create NB classifier using train data
87 Corpus.nb <- naiveBayes(Corpus.train.dtm.bin, Corpus.lab[Corpus.train], laplace=1)
88 #classify test data using the created NB classifier
89 Corpus.nb.pred <- predict(Corpus.nb, Corpus.test.dtm.bin)
90
91 #display a confusion matrix of the predictions versus the actual classes
92 table(Corpus.nb.pred, Corpus.lab[-Corpus.train])

```

**Figure 35. R Codes for Bag of Word Approach -Naive Bayes Technique – Continued**

As mentioned before, the data in BOC approach is in CSV format and already cleaned through the SKOS-Tool website by running queries and finding just most relevant concepts associated with specific class of interest. In BOC approach, four classification techniques are used. Naïve Bayes is used as baseline method. K-Nearest Neighbor (KNN), Random Forest, and Support Vector Machine (SVM) techniques are also used in order to improve the accuracy of the results. For two different scenarios, the same codes and techniques are used and the scenarios only changed the frequencies in the dataset. Following is the R code associated with Bag of Concept (BOC) approach.

Figure 36 shows the R code for the Naïve Bayes classifier in BOC approach.

```

1 library(e1071)
2 library(tm)
3 library(e1071)
4 library(wordcloud)
5
6 Data <- read.csv(file="C:/Corpus-Ramin/HeavyComplex.csv", stringsAsFactors <- TRUE)
7 str(Data)
8
9 #create training and test data
10 s <- sample(260, 200)
11 dataTrain1 <- Data[s, ]
12 dataTest1 <- Data[-s, ]
13
14 #create NB classifier using train data
15 XMLCorpus2.nb <- naiveBayes(XMLCorpus2.train.dtm.bin, Training.Data[,1], laplace=1)
16 #classify test data using the created NB classifier
17 XMLCorpus2.nb.pred <- predict(XMLCorpus2.nb,XMLCorpus2.test.dtm.bin)
18
19 #display a confusion matrix of the predictions versus the actual classes
20 table(XMLCorpus2.nb.pred,XMLCorpus2.lab[-XMLCorpus2.train])

```

**Figure 36. R Codes for Bag of Concept Approach – Naïve Bayes Techniques**

Figure 37 shows the R code for KNN method.

```

1 library(tm)
2 library(e1071)
3 library(XML)
4 library(wordcloud)
5 library(class)
6
7 Data <- read.csv(file="C:/Corpus-Ramin/HeavyComplex.csv", stringsAsFactors <- TRUE)
8 str(Data)
9 table(Data$class)
10
11 head(Data)
12 set.seed(9850)
13 gp <- runif(260)
14 gp
15 Data <- Data[order(gp) , ]
16 head(Data , 10)
17
18
19 summary(Data[, c(1,2,3,4,5,6,7,8)])
20 normalize <- function(x) {
21   return( (x--min(x))/ max(x)-min(x))
22 }
23
24 Data_n <- as.data.frame(lapply( Data[,c(1:24)], normalize))
25
26 Data_train <- Data_n[1:200, ]
27 Data_test <- Data_n[201:260, ]
28
29 Data_train_target <- Data[1:200,52]
30 Data_test_target <- Data[201:260,52]
31
32
33 m1 <- knn(train=Data_train , test=Data_test , cl=Data_train_target, k=7)
34 table(Data_test_target,m1)

```

**Figure 37. R Codes for Bag of Concept Approach - KNN Techniques**

Figure 38 shows the R code for Random Forest Technique.

```
1 library(randomForest)
2
3 Data <- read.csv(file="C:/HeavyComplex.csv", stringsAsFactors <- TRUE)
4
5 #get a sample of 200 out of 260 suppliers
6 s <- sample(260, 200)
7 dataTrain1 <- Data[s, ]
8 dataTest1 <- Data[-s, ]
9
10 rfm <- randomForest(Class ~ ., dataTrain1)
11 rfmPrediction <- predict(rfm, dataTest1)
12
13 table(dataTest1[,52], rfmPrediction)
14 mean(dataTest1[,52]==rfmPrediction)
15
16 # detecting the most important (distinguishing) concepts
17 importance(rfm)
18
19 getTree(rfm, 500, labelVar = TRUE)
```

**Figure 38. R Codes for Bag of Concept Approach - Random Forest Techniques**

Finally, the R code for SVM technique is shown in figure 39.

```
1 library(ggplot2)
2 library(e1071)
3
4 Data <- read.csv(file="C:/NewMethod.csv", stringsAsFactors <- TRUE)
5
6 s <- sample(260, 200)
7 Data_train <- Data[s, ]
8 Data_test <- Data[-s, ]
9
10 svmModel <- svm( Class ~., data= Data_train , kernel = "linear" , cost= .1, scale=FALSE)
11 print(svmModel)
12 plot(svmModel , Data_train)
13
14 p <- predict(svmModel, Data_test , type = "Class")
15 plot(p)
16 table(p, Data_test[, 52])
17 mean(p == Data_test[, 52])
```

**Figure 39. R Codes for Bag of Concept Approach - Random Forest Techniques**

## **APPENDIX C**

### **R Codes and Instructions for Unsupervised Learning Method**

In this section, the instructions and R codes is provided in order to conduct the unsupervised learning experiment. It consists of collecting and reading the data, applying some preprocessing functions, normalization of the frequencies, detecting the most distinguishing terms and documents and finally run the K-means Clustering and Topic Modeling algorithms. In the unsupervised learning experiment, the data is in .txt format and needs lots of preprocessing steps such as removing numbers, hyphens, symbols, punctuations, and some irrelevant and generic terms.

Here is the R code for the unsupervised learning method. Figure 40-45 show how to read and preprocess the data, how to find the most significant documents and N-grams.



```

1 library(ggplot2)
2 library(e1071)
3 library(caret)
4 library(quanteda)
5 library(irlba)
6 library(randomForest)
7
8 #ggplot2 for visualization
9 # caret for data partition
10 #quanteda is our main package in text analytics
11 #irlba singular variables decomposition
12
13 setwd("C:/Unsupervised/Corpus")
14
15 # Load up the .CSV data and explore in RStudio.
16 supplier.raw <- read.csv("HeavyMachine.csv", stringsAsFactors = FALSE)
17
18 # Clean up the data frame and view our handiwork.
19 supplier.raw <- supplier.raw[, 1:2]
20 names(supplier.raw) <- c("Label", "Text")
21
22 # Check data to see if there are missing values.
23 length(which(!complete.cases(supplier.raw)))
24
25 # Convert our class label into a factor.
26 supplier.raw$Label <- as.factor(supplier.raw$Label)
27
28 # The first step, explore the data.
29 # distribution of the class labels
30 prop.table(table(supplier.raw$Label))
31
32 # Visualize distribution with ggplot2, adding segmentation for ham/supplier.
33 library(ggplot2)
34
35 ggplot(supplier.raw, aes(x = TextLength, fill = Label)) +
36   theme_bw() +
37   geom_histogram(binwidth = 5) +
38   labs(y = "Text Count", x = "Length of Text",
39        title = "Distribution of Text Lengths with Class Labels")

```

**Figure 40. Unsupervised Learning – First Part**

```

42 # The quanteda package has many useful functions for quickly and
43 # easily working with text data.
44
45 library(quanteda)
46 help(package = "quanteda")
47
48 # Tokenize Supplier text
49 data <- tokens(train$Text, what = "word",
50               remove_numbers = TRUE, remove_punct = TRUE,
51               remove_symbols = TRUE, remove_hyphens = TRUE)
52
53 # Take a look at a specific supplier's text and see how it transfor
54 data[[21]]
55
56 # Lower case the tokens.
57 data <- tokens_tolower(data)
58
59 # Use quanteda's built-in stopword list for English.
60 data <- tokens_select(data, stopwords(),
61                       selection = "remove")
62 data[[21]]
63
64 # Create our first bag-of-words model.
65 data.dfm <- dfm(data, tolower = FALSE)
66
67 # Transform to a matrix and inspect.
68 data.matrix <- as.matrix(data.dfm)
69 View(data.matrix[1:20, 1:100])
70 dim(data.matrix)
71
72 # Investigate the effects of stemming.
73 colnames(data.matrix)[1:50]
74
75
76 # Setup a the feature data frame with labels.
77 data.df <- cbind(Label = train$Label, as.data.frame(data.dfm))
78 View(data.df)
79 #cbind function just bind two things together!
80 # Often, tokenization requires some additional pre-processing

```

**Figure 41. Unsupervised Learning – Second Part**

```

83 # Cleanup column names.
84 # if we have a bad name such as "8th" or "2nd" this function will fix them.
85 names(data.df) <- make.names(names(data.df))
86
87
88 # Use caret to create stratified folds for 10-fold cross validation repeated
89 # 3 times (i.e., create 30 random stratified samples)
90 set.seed(48743)
91 cv.folds <- createMultiFolds(train$Label, k = 3, times = 3)
92 cv.cntrol <- trainControl(method = "repeatedcv", number=10,
93                           repeats=3, index = cv.folds)
94 #we used index = cv.fold to mention our folds and
95 #do stratified class validation!
96
97 library(doSNOW)
98 # this function can run multiple train in parallel.
99 # it can run 3 times of 10 folds in same time.
100
101 # Time the code execution (we want to know the time of the run for this part)
102 start.time <- Sys.time()
103
104 # Create a cluster to work on 10 logical cores.
105 cl <- makeCluster(3, type = "SOCK")
106 registerDoSNOW(cl)
107 # each cluster number, uses a single core of the CPU!
108
109
110 rpart.cv.1 <- train(Label ~ ., data = data.df, method = "rpart",
111                    trControl = cv.cntrol, tuneLength = 7)
112 #method is the type of training. rpart is a single decision tree.
113
114 # Processing is done, stop cluster.
115 stopCluster(cl)
116
117 # Total time of execution on workstation was approximately 4 minutes.
118 total.time <- Sys.time() - start.time
119 total.time
120
121 # Check out our results.
122 rpart.cv.1

```

**Figure 42. Unsupervised Learning – Third Part**

```

123 # Our function for calculating relative term frequency (TF)
124 term.frequency <- function(row) { row / sum(row) }
125
126 # Our function for calculating inverse document frequency (IDF)
127 inverse.doc.freq <- function(col) {
128   corpus.size <- length(col)
129   doc.count <- length(which(col > 0)) log10(corpus.size / doc.count) }
130
131 # Our function for calculating TF-IDF.
132 tf.idf <- function(tf, idf) {tf * idf}
133
134 # First step, normalize all documents via TF.
135 data.df <- apply(data.matrix, 1, term.frequency)
136
137 # 1 means --> do it against the rows
138 # important --> this function transpose the data.matrix!
139
140 dim(data.df)
141 View(data.df[1:20, 1:20])
142
143 # Second step, calculate the IDF vector that we will use - both
144 # for training data and for test data!
145 data.idf <- apply(data.matrix, 2, inverse.doc.freq)
146 # 2 means --> apply this function on columns
147
148 # Lastly, calculate TF-IDF for our training corpus.
149 data.tfidf <- apply(data.df, 2, tf.idf, idf = data.idf)
150 dim(data.tfidf)
151 View(data.tfidf[1:50, 1:20])
152
153 # Transpose the matrix
154 data.tfidf <- t(data.tfidf)
155 dim(data.tfidf)
156 View(data.tfidf[1:20, 1:50])
157
158 # Check for incomplete cases.
159 incomplete.cases <- which(!complete.cases(data.tfidf))
160 train$Text[incomplete.cases]

```

**Figure 43. Unsupervised Learning – Forth Part**

```

162 # Fix incomplete cases
163 data.tfidf[incomplete.cases,] <- rep(0.0, ncol(data.tfidf))
164 dim(data.tfidf)
165 sum(which(!complete.cases(data.tfidf)))
166
167
168 # Make a clean data frame using the same process as before.
169 data.tfidf.df <- cbind(Label = train$Label, as.data.frame(data.tfidf))
170 names(data.tfidf.df) <- make.names(names(data.tfidf.df))
171 View(data.tfidf.df)
172
173
174 # Add bigrams and trigrams to our feature matrix.
175 data <- tokens_ngrams(data, n = 1:3)
176
177
178 # Transform to dfm and then a matrix.
179 data.dfm <- dfm(data, tolower = FALSE)
180 data.matrix <- as.matrix(data.dfm)
181 data.dfm
182 View(data.dfm)
183 dim(data.dfm)
184
185
186 # Normalize all documents via TF.
187 data.df <- apply(data.matrix, 1, term.frequency)
188
189
190 # Calculate the IDF vector that we will use for training and test data!
191 data.idf <- apply(data.matrix, 2, inverse.doc.freq)
192
193
194 # Calculate TF-IDF for our training corpus
195 data.tfidf <- apply(data.df, 2, tf.idf, idf = data.idf)

```

**Figure 44. Unsupervised Learning – Fifth Part**

```

198 # Transpose the matrix
199 data.tfidf <- t(data.tfidf)
200 View(data.tfidf)
201
202 # Fix incomplete cases
203 incomplete.cases <- which(!complete.cases(data.tfidf))
204 data.tfidf[incomplete.cases,] <- rep(0.0, ncol(data.tfidf))
205 sum(which(!complete.cases(data.tfidf)))
206
207
208 # Make a clean data frame.
209 data.tfidf.df <- cbind(Label = train$Label, as.data.frame(data.tfidf))
210 names(data.tfidf.df) <- make.names(names(data.tfidf.df))
211
212 #SVD
213
214 library(irlba)
215
216
217 # Perform SVD. Specifically, reduce dimensionality down to 300 columns
218 # for our latent semantic analysis (LSA).
219
220
221 #train.irlba <- irlba(t(data.tfidf), nv = 300, maxit = 600)
222 #I I will use the transposed of this matrix to have the same results as the video!
223
224 train.irlba <- irlba(data.tfidf, nu = 20 ,maxit = 20)
225
226
227 # nv = related to documents (right perspective)
228 # nu = related to terms (left perspective)
229 # maxit = most of the singular vector that we need to have in our raining data
230 #but we prefer 300!
231
232
233 # Take a look at the new feature data up close.
234 View(train.irlba$v)

```

**Figure 45. Unsupervised Learning – Sixth Part**

Figure 46 shows the clustering method which is applied for the output of the Term Frequency-Inverse Document Frequency method.

```
1 #DATA
2 Manufacturing <- read.csv("C:/Unseen.csv")
3
4
5 #REMOVE CLASS COLUMN
6 Manufacturing.features <- Manufacturing
7 Manufacturing.features$class <- NULL
8 View(Manufacturing.features)
9
10 #RUN THE KMEANS CLUSTERING
11 results <- kmeans(Manufacturing.features, 3)
12 results
13
14 #SIZE OF EACH CLUSTERS
15 results$size
16
17 #VECTOR OF THE CLUSTERS
18 results$cluster
19
20 results$totss
21 results$withinss
22 results$tot.withinss
23 results$betweenss
24 results$iter
25 results$ifault
26
27 #CLUSTERING PRECISION AND RECALL
28 table(Manufacturing$class , results$cluster)
29
30 #CLUSTERING PLOTS
31 plot(Manufacturing[c("Casting", "Machining")] , col = results$cluster)
32 plot(Manufacturing[c("Casting", "Machining")] , col = Manufacturing$class)
33 plot(Manufacturing[c("Complex.Machining", "Vertical.Boring")] , col = results$cluster)
```

**Figure 46. K-Means Clustering**

Finally, Figure 47 shows the topic modeling code.

```
244 library(topicmodels)
245
246 #Set parameters for Gibbs sampling
247 burnin <- 4000
248 iter <- 2000
249 thin <- 500
250 seed <- list(2003,5,63,100001,765)
251 nstart <- 5
252 best <- TRUE
253
254 #Number of topics
255 k <- 4
256
257 #Run LDA using Gibbs sampling
258 ldaOut <- LDA(data.dfm ,k, method="Gibbs",
259 control=list(nstart=nstart, seed=seed, best=best, burnin=burnin, iter=iter, thin=thin))
260
261 #write out results #docs to topics
262 ldaOut.topics <- as.matrix(topics(ldaOut))
263 write.csv(ldaOut.topics,file=paste("LDAGibbs",k,"DocsToTopics.csv"))
264
265 #top 10 terms in each topic
266 ldaOut.terms <- as.matrix(terms(ldaOut,20))
267 write.csv(ldaOut.terms,file=paste("LDAGibbs",k,"TopicsToTerms.csv"))
268
269 #probabilities associated with each topic assignment
270 topicProbabilities <- as.data.frame(ldaOut@gamma)
271 write.csv(topicProbabilities,file=paste("LDAGibbs",k,"TopicProbabilities.csv"))
272
273 #Find relative importance of top 2 topics
274 topic1ToTopic2 <- lapply(1:nrow(data.dfm),function(x)
275   sort(topicProbabilities[x,])[k]/sort(topicProbabilities[x,])[k-1])
276
277 #Find relative importance of second and third most important topics
278 topic2ToTopic3 <- lapply(1:nrow(data.dfm),function(x)
279   sort(topicProbabilities[x,])[k-1]/sort(topicProbabilities[x,])[k-2])
280
281 #write to file
282 write.csv(topic1ToTopic2,file=paste("LDAGibbs",k,"Topic1ToTopic2.csv"))
283 write.csv(topic2ToTopic3,file=paste("LDAGibbs",k,"Topic2ToTopic3.csv"))
```

**Figure 47. Topic Modeling Technique**



## REFERENCES

- Ameri, F., & Sabbagh, R. (2016). Digital Factories for Capability Modeling and Visualization. In *IFIP International Conference on Advances in Production Management Systems* (pp. 69–78).
- Bader, B. W., Berry, M. W., & Browne, M. (2008). Discussion tracking in Enron email using PARAFAC. In *Survey of Text Mining II* (pp. 147–163). Springer.
- Baker, R. P., & Maropoulos, P. G. (1998). Manufacturing capability measurement for cellular manufacturing systems. *International Journal of Production Research*, 36(9), 2511–2527.
- Barazandeh, B., Bastani, K., Rafieisakhaei, M., Kim, S., Kong, Z., & Nussbaum, M. A. (2017). Robust Sparse Representation-Based Classification Using Online Sensor Data for Monitoring Manual Material Handling Tasks. *IEEE Transactions on Automation Science and Engineering*.
- Bastani, K., Barazandeh, B., & Kong, Z. J. (2018). Fault Diagnosis in Multistation Assembly Systems Using Spatially Correlated Bayesian Learning Algorithm. *Journal of Manufacturing Science and Engineering*, 140(3), 31003.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.
- Chen, C. L. P., & LeClair, S. R. (1994). Integration of design and manufacturing: solving setup generation and feature sequencing using an unsupervised-learning approach. *Computer-Aided Design*, 26(1), 59–75.

- Chen, Y., & Lee, J. (2011). Autonomous mining for alarm correlation patterns based on time-shift similarity clustering in manufacturing system. In *Prognostics and Health Management (PHM), 2011 IEEE Conference on* (pp. 1–8).
- Cheraghi, S. H., Dadashzadeh, M., & Subramanian, M. (2004). Critical success factors for supplier selection: an update. *Journal of Applied Business Research*, 20(2), 91–108.
- Dong, B., & Liu, H. (2006). Enterprise website topic modeling and web resource search. In *Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on* (Vol. 3, pp. 56–61).
- 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* (pp. 193–199).
- Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: theory, algorithms, and applications* (Vol. 20). Siam.
- Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5), 403–420.
- Gupta, V., Lehal, G. S., & others. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, 1(1), 60–76.
- Harding, J. A., Shahbaz, M., Kusiak, A., & others. (2006). Data mining in manufacturing: a review. *Journal of Manufacturing Science and Engineering*, 128(4), 969–976.
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100–108.

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning* (pp. 485–585). Springer.
- Hayes, R. H., & Wheelwright, S. C. (1984). *Restoring our competitive edge: competing through manufacturing* (Vol. 8). John Wiley & Sons New York, NY.
- Huang, L., & Murphey, Y. L. (2006). Text mining with application to engineering diagnostics. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 1309–1317).
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3), 264–323.
- James, M. (1985). *Classification algorithms*. Wiley-Interscience.
- Jiao, J. R., Zhang, L. L., Pokharel, S., & He, Z. (2007). Identifying generic routings for product families based on text mining and tree matching. *Decision Support Systems*, 43(3), 866–883.
- Jiao, J., Zhang, L., Zhang, Y., & Pokharel, S. (2008). Association rule mining for product and process variety mapping. *International Journal of Computer Integrated Manufacturing*, 21(1), 111–124.
- Jung, K., Kulvatunyoo, B., Choi, S., & Brundage, M. P. (2016). An overview of a smart manufacturing system readiness assessment. In *IFIP International Conference on Advances in Production Management Systems* (pp. 705–712).
- Kaplan, R. M. (2005). A method for tokenizing text. *Inquiries into Words, Constraints and Contexts*, 55.
- Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, 1(6), 90–95.

- Kuhn, M., & others. (2008). The caret package. *Journal of Statistical Software*, 28(5), 1–26.
- Landauer, T. K. (2006). *Latent semantic analysis*. Wiley Online Library.
- Lee, J., & Hong, Y. S. (2016). Extraction and visualization of industrial service portfolios by text mining of 10-K annual reports. *Flexible Services and Manufacturing Journal*, 28(4), 551–574.
- Lekurwale, R. R., Akarte, M. M., & Raut, D. N. (2015). Framework to evaluate manufacturing capability using analytical hierarchy process. *The International Journal of Advanced Manufacturing Technology*, 76(1–4), 565–576.
- Lin, T. Y., Xiao, Y., Yang, C., Liu, X., Li, B. H., Guo, L., & Xing, C. (2016). Manufacturing Capability Service Modeling, Management and Evaluation for Matching Supply and Demand in Cloud Manufacturing. In *Theory, Methodology, Tools and Applications for Modeling and Simulation of Complex Systems* (pp. 35–48). Springer.
- Liu, C., Jiang, P., & Cao, W. (2014). Manufacturing capability match and evaluation for outsourcing decision-making in one-of-a-kind production. In *Computer Supported Cooperative Work in Design (CSCWD), Proceedings of the 2014 IEEE 18th International Conference on* (pp. 575–580).
- Luo, Y., Zhang, L., Tao, F., Ren, L., Liu, Y., & Zhang, Z. (2013). A modeling and description method of multidimensional information for manufacturing capability in cloud manufacturing system. *The International Journal of Advanced Manufacturing Technology*, 69(5–8), 961–975.

- Malakooti, B., & Yang, Z. (1995). A variable-parameter unsupervised learning clustering neural network approach with application to machine-part group formation. *International Journal of Production Research*, 33(9), 2395–2413.
- Menon, R., Tong, L. H., & Sathiyakeerthi, S. (2005). Analyzing textual databases using data mining to enable fast product development processes. *Reliability Engineering & System Safety*, 88(2), 171–180.
- Menon, R., Tong, L. H., Sathiyakeerthi, S., Brombacher, A., & Leong, C. (2004). The needs and benefits of applying textual data mining within the product development process. *Quality and Reliability Engineering International*, 20(1), 1–15.
- Miles, A., & Bechhofer, S. (2009). SKOS simple knowledge organization system reference. *W3C Recommendation*, 18, W3C.
- Miltenburg, J. (2005). *Manufacturing strategy: how to formulate and implement a winning plan*. CRC Press.
- Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*, 18.
- Nigel, S., & Michael, L. (2008). *Operations Strategy*. Harlow: Prentice Hall Financial.
- Ramos, J., & others. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, pp. 133–142).
- Romanowski, C. J., & Nagi, R. (2002). A data mining and graph theoretic approach to building generic bills of materials. In *IIE Annual Conference. Proceedings* (p. 1).
- Romanowski, C. J., & Nagi, R. (2004). A data mining approach to forming generic bills of materials in support of variant design activities. *Journal of Computing and Information Science in Engineering*, 4(4), 316–328.

- Sabbagh, R., & Ameri, F. (2017). A Thesaurus-Guided Text Analytics Technique for Capability-Based Classification of Manufacturing Suppliers. In *ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (p. V001T02A075--V001T02A075).
- Sabbagh, R., Ameri, F., & Yoder, R. (2018). Thesaurus-guided Text Analytics Technique for capability-based classification of manufacturing Suppliers. *Journal of Computing and Information Science in Engineering*, 18(2).  
<https://doi.org/10.1115/1.4039553>
- Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507–2517.
- Salami, S. R., TaghaviFard, M. T., & Majidifar, M. M. (2015). Identifying the Factors of Technological Capability Assessment--Case study Pressing parts Manufacturing Department of Iran Khodro.
- Shotorbani, P. Y., Ameri, F., Kulvatunyou, B., & Ivezic, N. (2016). A Hybrid Method for Manufacturing Text Mining Based on Document Clustering and Topic Modeling Techniques. In *IFIP International Conference on Advances in Production Management Systems* (pp. 777–786).
- Skinner, W. (1969). Manufacturing-missing link in corporate strategy.
- Srivastava, A. N., & Sahami, M. (2009). *Text mining: Classification, clustering, and applications*. Chapman and Hall/CRC.
- Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., ... He, X. (2014). Interpreting the public sentiment variations on twitter. *IEEE Transactions on Knowledge and Data Engineering*, 26(5), 1158–1170.

- Teece, D. (1990). Firm capabilities, resources and the concept of strategy. *Economic Analysis and Policy*.
- Torkul, O., Cedimoglu, I. H., & Geyik, A. K. (2006). An application of fuzzy clustering to manufacturing cell design. *Journal of Intelligent & Fuzzy Systems*, 17(2), 173–181.
- Ur-Rahman, N., & Harding, J. A. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications*, 39(5), 4729–4739.
- Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning* (pp. 977–984).
- Wang, F., Wang, Z., Li, Z., & Wen, J.-R. (2014). Concept-based short text classification and ranking. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 1069–1078).
- Wang, K. (2007). Applying data mining to manufacturing: the nature and implications. *Journal of Intelligent Manufacturing*, 18(4), 487–495.
- Xue, D., & Dong, Z. (1997). Coding and clustering of design and manufacturing features for concurrent design. *Computers in Industry*, 34(1), 139–153.
- Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1445–1456).
- Zhai, Z., Liu, B., Xu, H., & Jia, P. (2011). Constrained LDA for grouping product features in opinion mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 448–459).

- Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. In *ACM Sigmod Record* (Vol. 25, pp. 103–114).
- Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF\* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, 38(3), 2758–2765.