TEXT MINING TECHNIQUES FOR ANALYZING UNSTRUCTURED MANUFACTURING DATA

by

Peyman Yazdizadeh Shotorbani

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
August 2016

Committee Members:

Farhad Ameri, Chair

Jaymeen Shah

Vedaraman Sriraman

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

DEDICATION

I would like to dedicate this thesis work to my wonderful wife, Mahsa, who has been a tremendous source of support and encouragement during the challenges of graduate school and life. I am truly grateful for having you in my life. This work is also dedicated to my beloved parents, Arman Yazdizadeh and Latifeh Torabi who have always loved me unconditionally. They have been always there to make sure I am doing well even though we are thousands of miles away. I would also like to extend my dedication to my beloved sister, Parisa who always supports his little brother and wants me always the bests.

ACKNOWLEDGEMENTS

This thesis work was accomplished at the Engineering Informatics Research Group Lab (INFONEER), Department of Engineering Technology at Texas State University, under direction of Dr. Farhad Ameri. This thesis is partially funded by NIST cooperative agreement with Texas State University No. 70NANB14H255.

A very special thanks to Dr. Farhad Ameri for his countless hours of reflecting, reading, encouraging, and most of all patience throughout the entire process of my graduate life. He inspired and supported me by all means he could. I really appreciate the opportunity of work under his supervision in INFONEER Lab.

The members of my thesis committee, Dr. Sriraman and Dr. Shah have generously given their valuable time and expertise to better my work. I thank them for their contribution and their constant supports.

TABLE OF CONTENTS

Page

LIST OF TABLES

Tables **Page**

LIST OF FIGURES

Figures **Page**

LIST OF ABBREVIATIONS

Abbreviation	Description
AM	Additive Manufacturing
DTM	Document Term Matrix
LDA	Latent Dirichlit Allocation
NLP	Natural Language Processing
PLSA	Probabilistic Latent Semantic analysis
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language
SSE	Sum of Squared Error
SVM	Support Vector Machine
TM	Text Mining
XML	EXtensible Markup Language

ABSTRACT

Manufacturing companies are increasingly enhancing their web presence as a strategy for improving their visibility in the global market. The exponential growth of manufacturing websites has resulted in a drastic increase in the size and variety of unstructured manufacturing information available online. This poses both challenges and opportunities. The challenge is related to efficient information search and retrieval when dealing with a large volume of heterogeneous information. Traditional search methods, such as keyword search, can no longer meet the information retrieval and organization needs of the manufacturing cyberspace. At the same time, the textual data available online contains a wealth of technical knowledge that, if mined properly, may result in discovery of new patterns and trends that were otherwise unknown. There is an acute need for more advanced computational tools and techniques that can help search, organize , and summarize large achieves of text pertaining to technological capabilities of manufacturing suppliers . In this research, three text mining techniques, namely, Classification, Clustering, and Topic Modeling are applied to analyzing manufacturing data. R programming package is used for implementation of the aforementioned techniques. The novelty of the proposed classification technique is in adopting concept-based method rather than term-based method that results in higher semantic relevance of the results. Also, clustering and topic modeling are serialized to improve the likelihood of discovering useful knowledge patterns.

I. INTRODUCTION AND MOTIVATIONS

Introduction

According to recent studies, 90% of the data in digital space will be unstructured in upcoming decade (Gantz & Reinsel, 2011). Unstructured data are heavily loaded with texts and they can be found almost everywhere such as in electronic documents, online webpages, and social media. Natural language text is readable and understandable for human. However, it is not feasible for human to process high volume of textual data and classify and retrieve patterns, sentiments or meanings out of the data. Text analytics can help classify and organize the data and extract rules and patterns from unstructured textual data. Computerized text analytics approaches are currently being implemented in diverse industries for discovering and predicting trends and knowledge patterns in textual data (Chakraborty, Pagolu & Garla, 2013). For instance, text analytics can use upstream textual data of social media for crime prediction and prevention (Gerber, 2014). It can also be used to process financial statements to detect possible frauds cases (Singh, 2012). Pharmaceutical industries are also taking advantage of text analytics to mine biomedical documents to discover more beneficial drugs (Ku, Chiu, Zhang, Chen & Su, 2014; Loging, n.d.).

Text analytics utilizes different techniques and tools to derive insights from unstructured data. These techniques, based on their functionalities can generally be categorized as classification, information retrieval (IR), concept mining, summarization, exploratory analysis, ontology management, etc. From above-mentioned techniques, classification, summarization and exploratory analysis are in the realm of *Text Mining*. Exploratory analysis comprises *topic extraction* and *cluster analysis*. Therefore, text mining can be

regarded as a subset of text analytics which emphasizes on data mining procedures by using Natural Language Processing (NLP) and machine learning techniques (Chakraborty et al., 2013). In this research, text mining techniques are applied to manufacturing supplier classification and discovery problem. This work is among the first research efforts that use text mining techniques for problem solving in manufacturing field.

Supplier Discovery Problem

One of the main difficulties in rapid configuration of virtual supply chains in a distributed setting, is finding the right suppliers who possess the required set of technological capabilities and competencies. The relationship between the participants of distributed supply chains, particularly at the early stages of supply chain formation, is often virtual and based on electronic and web-based interactions. This presents a challenge to effective supplier discovery and evaluation.

There are two approaches for supplier discovery in virtual environments, centralized approach and decentralized approach. The centralized approach for supplier discovery utilizes centralized databases managed by web portals, such as Thomas Net[1] and MFG.com[2] that provide single points of reference for both customers and suppliers. These web portals offer different search tools and mechanisms with varying levels of complexity and precision. Due to the well-structured information model used in the backend database of such portals, the search process is computationally efficient and inexpensive. However, these portals often fail to accurately connect suppliers and customers for multiple reasons. First, their search process is typically keyword-based which uses a limited array of categories and criteria for characterizing suppliers and

[1] http://www.thomasnet.com
[2] http://www.mfg.com/

hence, suppliers are not properly differentiated. Second, their underlying information models are static in nature and do not evolve in a timely fashion to reflect the technological changes in the manufacturing industry. Additionally, the rigid templates that suppliers have to use for describing their capabilities do not provide them with enough flexibility and expressivity to freely describe their capabilities.

The decentralized approach for supplier search is based on direct web search using generic search engines. The online profiles of manufacturing suppliers, directly maintained on the firms' websites, contain a wealth of information pertaining to the capabilities of suppliers represented in natural language. In their webpages, manufacturing companies typically provide different types of information such as their primary and secondary services, their machineries, the materials they can process, the processes they are expert in, and the types of products or geometries they can produce. The contents of suppliers' profiles are not constrained by a particular vocabulary imposed by the online portals. Therefore, they reflect the true capabilities of manufacturers more accurately and realistically. Also, the content of online profiles are updated more frequently compared to the profiles on the centralized search portals. However, the unstructured and heterogeneous nature of the contents of online profiles is an impediment to an efficient supplier discovery experience. Furthermore, the sheer size of the data available online makes the decentralized search process a daunting task. That is why basic Google search usually cannot provide satisfactory results.

To improve the intelligence and precision of the decentralized search process, the search tools should be customized and specialized for supplier search purpose but the data should remain within the control of suppliers. To this end, it is first necessary to

narrow down the scope of the search through classifying and clustering suppliers into various categories based on their technical similarities. For example, by defining a subset of suppliers that are expert in providing precision machining services for medical industry, the interested user will deal with a group of highly relevant suppliers that can be further drilled down and explored. After classification, customized queries can be formulated over each family of suppliers to improve the relevance of the returned set.

Problem statement

Currently about 90 percent of available data are in form of plain texts known as unstructured data. Unstructured text mining rapidly is finding vital role for diverse business applications. Applying different text mining tools and approaches enable customers to easily find manufacturing suppliers who meet their business requirements. This makes the supplier discovery experience much faster and less risky than traditional styles and therefore, it reduces extra business costs for both parties. So in other words, text mining approaches can be applied in manufacturing industries to answer questions related to the supply chain configurations to enhance daily operational efficiencies along with improving strategic decisions. *"The objective of this research work is to develop and implement diverse text analytics techniques to automatically discover and build supplier families based on the textual capability narratives available in their online profiles. This research work applies text mining, NLP and machine learning techniques to intelligently and without human intervention to classify and summarize suppliers based on diverse process services they offer".*

Research Questions

This thesis work is trying to answer to the following questions.

- How supervised text classification techniques can help rapid configuration of supply chain in manufacturing domain?
- How performance of the text mining techniques in manufacturing domain can be improved?
- How Clustering and Topic modeling add extra values to rapid supply chain configuration?

Research Methodology

Methodology: As mentioned earlier, text mining includes diverse techniques such as document classification, document clustering, associations and topic modeling. These different text mining techniques, which can be developed based on diverse mathematical and probabilistic algorithms, can be applied to different problems depending on their specific needs. The advantages and disadvantages of different techniques and their corresponding algorithms are investigated first in order to identify the most appropriate techniques for the supplier classification and characterization problem. After the best method is selected, data gathering process will be dealt with. Among different industrial processes in manufacturing domain, Casting and Machining processes are selected as examples of simple dissimilar classes, and also Turning and Milling processes are selected as instances for simple similar classes. Data related to these classes are gathered from websites of manufacturing suppliers and are kept as documents in a corpus. Before any text mining techniques can be applied in a corpus, diverse cleaning processes is needed to be executed to the documents of the corpus to mostly retrieve useful and

technical terms. Figure 1 illustrates an example of a manufacturing related document after cleaning processes.

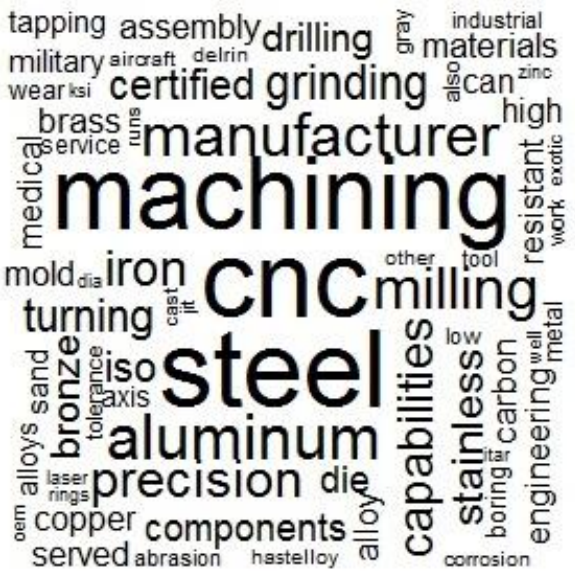


Figure 1. Corpus after refinement

The cleanup process, or preprocessing of the corpus, is composed of multiple steps such as removal of English stop words, removal of the generic words and removal of punctuations, numbers, and whitespaces. The generic words are gathered and removed manually. The font size of the terms in Figure 1 indicates the degree of importance of the terms in the corpus based on their frequency of occurrence.

After this preprocessing stage, text classification, clustering, association and topic modeling methods can be applied on the corpus. Based on the literature review, best algorithm for text classification, clustering and topic modeling are respectively known as Naïve Bayes, K-Means and Latent Dirichlet Allocation (LDA). R programming language is used to design and implement applications for these techniques.

Document clustering, association and topic modeling are unsupervised subtasks of text mining where text/document classification is a supervised subtask of text mining.

Therefore, for text classification technique, there is a need to have a valid and high quality training data so the developed classifier can automatically predict and classify any given test data with high precision (websites of manufacturing suppliers) under their relevant categories. In fact the performance of the classifier not only depends on the quantity but also quality of the prepared training data. In this thesis work, training data for classification task will be obtained through two different sources. The first source is the corpus which keeps preprocessed documents and the second source is a manually crafted thesaurus (ManuTerms) for manufacturing concepts. Finally, impact of each training data on the performance of the classifier will be investigated.

Research Tasks

Table 1 shows the work details which are going to be included in this research. Nine different tasks need to be accomplished during this research work.

Table 1. Research Tasks

Research Question 1	How supervised text classification techniques can help rapid configuration of supply chain in manufacturing domain?
Related Tasks	1. Literature reviews for being familiar with the state of the art techniques used in text mining. 2. Gathering proper data through web surfing. 3. Preprocessing data. 4. Developing R-based programs for text classifications.

Research Question 2	How performance of the text mining techniques in manufacturing domain can be improved?
Related Tasks	5. Investigating advantage of SKOS and ManuTerms as an alternative for training data used in the R-based programs. 6. Designing a set of required semantic queries to meet requirements of task 5. 7. Design of additional filters in preprocessing steps 8. Experimental validation
Research Question 3	How clustering and association rules can add extra values to supply chain configuration?
Related Tasks	9. Developing and implementing R-based programs for text clustering and associations.
Research Question 4	How hidden patterns in manufacturing documents can be revealed?
Related Tasks	10. Developing and implementing R-based Topic Modeling to uncover hidden themes in manufacturing documents

Assumptions

A general assumption for this thesis work is as follow:

- Each test document in classification task only belongs to a predefined class.

Limitation

- Homepages of manufacturing suppliers (known as documents) usually are providing information in different processes, this make it difficult to make a training data or a data set for a single process.
- There is not a process-based manufacturing data set as a benchmark, therefore this research uses its own designed data set.

Delimitation

Websites of manufacturing suppliers are made of multiple layers of connected pages. To have a short but comprehensive sample data which can represent capability of their suppliers, only the relevant pages, including the homepage, are considered as the sample data of the experiment.

II. A SUPERVISED TEXT MINING TECHNIQUE FOR CLASSIFICATION OF MANUFACTURING SERVICE PROVIDERS

Introduction

Classification is a data mining technique used for predicting group membership for instances of data (Agarwal, Thakare & Jaiswal, 2015). In the manufacturing domain, classification is used in different applications such as fault diagnosis, quality control, and condition monitoring (Martens & Provost, 2014). Text Classification, as a subset of classification techniques, is the task of classifying documents by their content under predefined categories. With exponential growth in the volume of the unstructured data in the Internet, automated text classification has become increasingly important as it helps categorize and organize various heterogeneous documents into different classes of interest with known properties, thus making information search and retrieval much more efficient. In this chapter, text classification problem is investigated in a manufacturing setting. In particular, this chapter deals with classification of suppliers of manufacturing services based on their capability narratives extracted from their online profiles.

Supplier Classification Problem

Supplier classification is one of the necessary steps in supply chain management. It is conducted either for supplier evaluation and selection or for deciding the strategies to be followed after supplier selection. By classifying manufacturing suppliers based on different criteria such as their core services, capabilities, constraints, and target customers and industries, the supplier discovery process is streamlined through narrowing down the search space to only highly relevant suppliers. For example, by creating a class of service providers that can perform cost estimation for additive manufacturing processes, more advanced and customized queries can be run against the selected subset in order to find

more specific providers that, for instance, focus on small metallic parts for aerospace industry. This will improve both the visibility of the service provider and the computational efficacy of the service discovery process.

Most of the existing product sourcing and supplier discovery platforms provide some type of supplier classification based on the services and products they provide. However, supplier classes, or categories, are assigned explicitly by the suppliers or the portal administrators at the time of registration. Table 2 shows the subcategories of CNC machining in Thomas Net platform together with the number of suppliers under each category. Under the implemented classification scheme in Thomas Net platform, one supplier can belong to multiple categories.

Table 2. The sub-categories of CNC Machining category in Thomas Net

1. Metal & Wood CNC Patterns (38 suppliers)	10. Fiber Optic Component CNC Machining (88 suppliers)
2. CNC Plastics Routing (178 suppliers)	11. CNC Glass Machining (122 suppliers)
3. General CNC Machining (10,697 suppliers)	12. Large CNC Machining (648 suppliers)
4. OEM Parts CNC Machining (731 suppliers)	13. Medical CNC Machining (1,194 suppliers)
5. CNC Swiss Machining (323 suppliers)	14. Short Run CNC Machining (1,267 suppliers)
6. 9-Axis CNC Machining (103 suppliers)	15. Ultra Precision 5-Axis CNC Machining (287 suppliers)
7. Aircraft & Aerospace CNC Machining (1,166 suppliers)	16. CNC Punching (745 suppliers)
8. Castings CNC Machining (363 suppliers)	17. CNC Turret Punching (219 suppliers)
9. Exotic Metal CNC Machining (890 suppliers)	

Manual and explicit classification is a reasonable approach in centralized scenarios where there exists a central repository of suppliers and the input data is homogenized and unified during the registration process using templates or reference vocabularies. However, for decentralized scenarios in which the search space is extended to the entire

Internet and there is no unifying template available, a different approach to classification should be adopted.

Data or Meta-data?

Contract manufacturing suppliers can be classified based on annotations, or meta-data, grounded in some standardized reference model. For example, they can be manually tagged by some labels coming from a reference vocabulary and classified based on the tags they have received. Thomas Net classification is an example of classification based on meta-data since supplier profiles are tagged with pre-defined service categories when they are created. Annotation-based classification works well only if the vocabulary in the domain of discourse is more or less stable and standardized. However, in a dynamic environment where the terminology is constantly evolving, classification based on meta-data will not always produce accurate results. It is because reference model evolution and adaptation typically lags behind changes in the domain knowledge especially when the model change management process undergoes an approval workflow that involves multiple stakeholders. This delayed alignment would adversely impact the performance of annotation-based classifiers. The alternative solution is to directly use supplier capability narratives, in the form of natural language, as the source of information for service classification. Textual resources evolve faster than their associated ontologies and reference models and, therefore, they more accurately reflect the evolution of the domain itself (Maynard, Peters, d'Aquin & Sabou, 2007).

Research Objective and Approach

The primary objective of the research reported in this chapter is to use *text classification* techniques to automatically characterize and classify manufacturing

suppliers based on their capability narratives available in their online profiles. Each profile is treated as a *document* to be classified under predefined classes. Naïve Bayes is used as the underlying algorithm for the proposed classification method. Simplicity, scalability, and its processing speed are among the advantages of Naïve Bayes algorithm (Ting, Ip, & Tsang, 2011). 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 (Ting et al., 2011). Like most supervised text classification algorithms, Naïve Bayes classifier requires training data. The performance of text classifiers very much depends on the size and quality of the training data. This work is not intended to alter or improve the underlying mathematical algorithm of the Naïve Bayes classifier. Instead, the focus is on devising a suitable approach for training data extraction, preparation, and representation for supplier classification problem and also improving the semantic relevance of the results. The specific questions that are explored in this chapter include:

- *Is text mining a viable method for supplier classification?*
- *What is best source of training data for supplier classification using text mining techniques?*
- *How to reduce the cost of training data preparation?*
- *How to pre-process the training data for more efficient machine learning?*
- *How to prove the accuracy and semantic relevance of the text classifier?*

To create a baseline, a standard term-based text classifier is developed first in this work. A set of pre-classified documents directly extracted from online supplier profiles is used to train the term-based text classifier. The second developed classifier is a concept-

based classifier. While terms are regarded as *lexical entities*, concepts are considered to be *semantic entities*. Since the concept-based classifier performs classification in the conceptual space, its resulting classes are expected to be more consistent semantically. For the concept-based classifier, the training data is obtained from ManuTerms. ManuTerms is a hand-made thesaurus of manufacturing concepts (Ameri, Kulvatunyou, Ivezic & Kaikhah, 2014). The novelty of the proposed thesaurus is that it uses Simple Knowledge Organization Systems (SKOS) for thesaurus representation ("SKOS Simple Knowledge Organization System Reference", 2016). Since ManuTerms is connected to the Linked Open Data (LOD), it can be extended and validated by the community of users in a collaborative fashion. Therefore, the development and evolution cost is significantly lower than that of isolated thesauri created by teams of domain experts in a top-down manner. The hypothesis that is tested in this work is that concept-based approach improves the performance of the classifier with respect to precision and recall due to its semantic nature. This hypothesis is tested experimentally using standard information retrieval metrics such as precision, recall, and F-measure.

Related Work

Classification is a learning process that maps a data point into one of several predefined classes. There are two broad categories of text classification techniques, namely, *single-label classification* and *multi-label classification*. In the single-label classification technique, a document is classified under one class while in the multi-label classification, documents can belong to multiple classes. Text classification methods use several tools and techniques from information retrieval and machine learning. Some of the standard methods for text classification include Naive Bayes, linear and nonlinear

Support Vector Machines (SVMs), tree-based classification, neural network and K-nearest neighbors (Martens & Provost, 2014; Sebastiani, 2002). These methods can be for text classification in the framework of terms extraction (Bijalwan, Kumar, Kumari & Pascual, 2014; Frigui & Nasraoui, 2004). Research by Steinbach, Karypis and Kumar (2000) provides a comparison between different classifications methods. Research works by Korde (2012) and Aggrawal and Yu (2009) investigate the drawback of different text mining methods in the context of pattern building. These methods are used in different applications such as sentiment analysis (Pang & Lee, 2008), spam detection (Lau et al., 2011), web page classification (Qi & Davison, 2009), knowledge management in medical informatics (Chen, Fuller, Friedman & Hersh, n.d.), text classification for topical web crawl (Pant & Srinivasan, 2005) and in construction industry for mining and categorizing Post Project Reviews (Ur-Rahman & Harding, 2012).

Researches by Sanchez-Pi, Martí and Garcia (2014) and also Bérenguer, Grall and Soares (2012) investigate applications of text mining techniques respectively in oil and aerospace industries. These studies especially focus on safety assessment procedures. The main goal of these researches are assigning arriving safety event reports to a set of predefined categories based on their textual context.

Text mining has also application in engineering diagnostics. For instance, in automotive industry, the descriptions of auto problem are often used manually for fault diagnostic process in a way that problem should be assigned to its predefined diagnostic categories such as engine, transmission and etc. This procedures for auto fault diagnosis can be replaced by an automatic text document categorization (Lu Murphey, 2015). However, with the advance of technology in auto industry, many more diagnosis

categories are being generated and due to the nature of data preparation which is mainly based on field surveys, the presented technique needs frequent maintenance.

Choudhary, Harding and Tiwari (2008) reviewed the existing works in the manufacturing domain with focus on knowledge discovery based on data mining techniques such as association, classification, clustering and evolution analysis. They also claimed that application of data mining in the context of manufacturing processes and enterprises is growing rapidly. However, Choudhary et al. (2008) just conducted a general study and they did not go through any specific process.

Applying text mining in manufacturing domain is in its early stages and there are few works reported in this field. For instance, Kornfein and Goldfarb (2007) introduced one such application for manufacturing quality defects and service shop datasets. According to their study, service reports such as discoveries during a repair or documentation of manufacturing quality problem, which are considered as text data, are created by technician and contain important information about system design. These data need to be automatically classified to be used by domain experts for further analysis. This study claimed that the SVM method is one of the best fit for classification of technical passages and result of their research work can be applied for other domains (M. Kornfein & Goldfarb, 2007). Kung, Lin and Hsu (2015) investigated the application of text classification techniques for identifying quality issues in semiconductor manufacturing based on the unstructured data available in hold records. As another example, Liu, Loh, Toumi and Tor (2008) proposed a text mining technique for building a common corpus for manufacturing information retrieval. However, they did not suggest

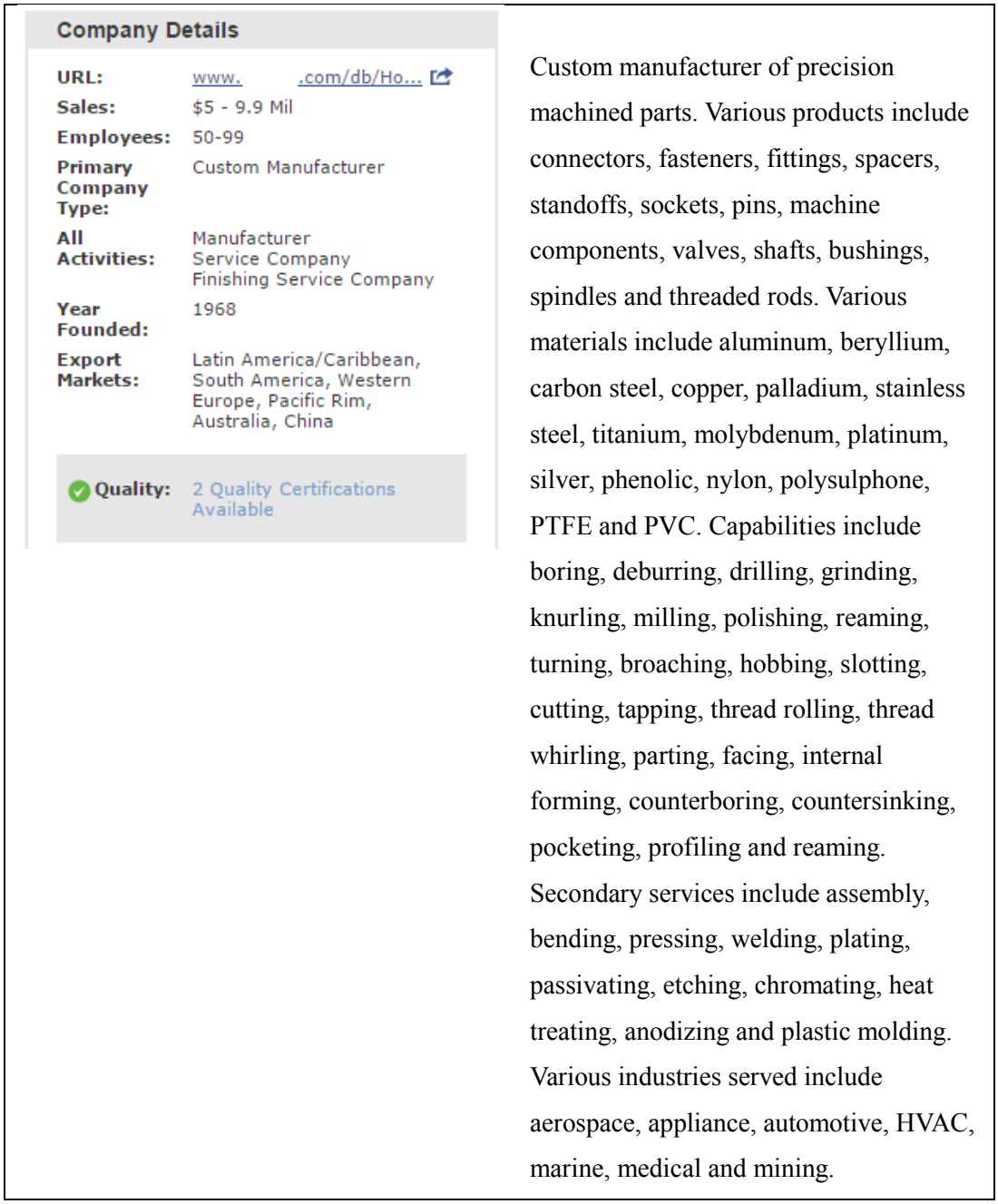
concrete applications for the proposed corpus. This chapter focuses on manufacturing service classification based on the textual description of their capabilities.

Manufacturing Service Description

Service can be defined as the application of competences for the benefits of others (Spohrer, Maglio, Bailey & Gruhl, 2007). In the context of manufacturing industry, a service typically refers to a *contract manufacturing service* in which the manufacturing service provider (MSP) utilizes machinery, labor, and energy to transform material inputs to finished or semi-finished goods according to the contract specifications. There are also some services that are more computational and intangible, rather than physical, in nature such as a service that estimates the assembly time for an assembly. We refer to this group of services as *computational manufacturing services*. The focus of this work, however, is on contract manufacturing services.

The descriptions of contract manufacturing services can be found either on the sourcing portals or on the websites of manufacturing suppliers. Sourcing portals, such as Thomas Net and MFG.com, provide templates for both service consumers and providers such that they can describe their capabilities and needs and find the right manufacturing counterparts. In their profiles, suppliers can provide a capability narrative, typically up to 200 words, that summarizes their technological capabilities in terms of processes, services, materials, products, and industries. Figure 2 shows an example capability narrative obtained from Thomas Net. *Portal narratives* are more or less homogenous in terms of text structure, size, content, and vocabulary. Also, they are rather dense with respect the concentration of keywords with manufacturing relevance. Therefore, they are

suitable resources for building and expanding a corpus of manufacturing terms which is needed for text classification purpose.



Company Details

URL:	www. .com/db/Ho...
Sales:	$5 - 9.9 Mil
Employees:	50-99
Primary Company Type:	Custom Manufacturer
All Activities:	Manufacturer Service Company Finishing Service Company
Year Founded:	1968
Export Markets:	Latin America/Caribbean, South America, Western Europe, Pacific Rim, Australia, China
Quality:	2 Quality Certifications Available

Custom manufacturer of precision machined parts. Various products include connectors, fasteners, fittings, spacers, standoffs, sockets, pins, machine components, valves, shafts, bushings, spindles and threaded rods. Various materials include aluminum, beryllium, carbon steel, copper, palladium, stainless steel, titanium, molybdenum, platinum, silver, phenolic, nylon, polysulphone, PTFE and PVC. Capabilities include boring, deburring, drilling, grinding, knurling, milling, polishing, reaming, turning, broaching, hobbing, slotting, cutting, tapping, thread rolling, thread whirling, parting, facing, internal forming, counterboring, countersinking, pocketing, profiling and reaming. Secondary services include assembly, bending, pressing, welding, plating, passivating, etching, chromating, heat treating, anodizing and plastic molding. Various industries served include aerospace, appliance, automotive, HVAC, marine, medical and mining.

Figure 2. Capability narrative of a machining service provider obtained from Thomas Net (portal narrative)

The second source of manufacturing service narratives is the websites of manufacturing suppliers (*website narrative*). The online profiles of manufacturing service

providers, directly maintained on the firms' websites, contain a wealth of information pertaining to suppliers' capabilities and their services. In their webpages, manufacturing companies typically provide information regarding their primary and secondary services, their machineries and equipment, the materials they can process, and the types of products or geometries they can produce. Figure 3 shows the capability narrative excerpted from the same supplier described in Figure 2. The content of suppliers' profiles is not constrained by any particular vocabulary imposed by a reference model or template. Therefore, they reflect the true capabilities of manufacturers more accurately and realistically. One challenging aspect of text classification based on the website narratives is variation in the length of the text as well as the density and usefulness of information. Also, the vocabulary used for service description varies from one supplier to another. Therefore, text classification methods that merely rely on term frequency, and ignore the semantics, would not yield reliable results. This is a hypothesis that is investigated in this research.

In this chapter, both portal narratives and website narratives are used as test and training data.

For over 68 years **XYZ Parts, Inc.** has been crafting precision-machined parts for North American companies, becoming an integral part of the supply chain of many industry-leading manufacturers. Our employees have a real passion for manufacturing, for getting down to the nitty-gritty of making complex machined parts that are designed to perform critical functions in our customer's products. We see ourselves more as partners with our customers, trying to find new ways to make their parts perform better, last longer and cost less. By doing this we help our

customers win a greater market share which, in turn, results in more business for all of us! The design, manufacture and assembly of custom machined parts is our only business and is one that we are absolutely committed to doing well, consistently for all of our customers. We never "take our eye off the ball" when it comes to listening to our customer and making absolutely sure that we are giving them 100% service and attention to their needs. At XYZ, we utilize both conventional screw machines as well as state-of-art multi-axis CNC turning equipment so that we can guarantee our customers that they are getting the best part for their money, made using the most efficient process. We can machine virtually all metals and plastics in sizes ranging from .0625" up to 3 ½" in diameter. We can efficiently process low volume orders using generic tooling on CNC lathes or we can tool your job to run on custom-designed machinery that is among the fastest in the world. When a part cannot be completed in one operation on our machines, our high speed vertical machining centers and CNC chuckers can finish your part with only a minimum of additional set up time needed.

Figure 3. Capability narrative directly obtained from manufacturers websites (website narrative)

Classification Using Term-Based Training Data

To create a baseline, supplier narratives are classified using the term-based method that is described in detail in this section. This experiment also partially addresses the first research question regarding the viability of text mining approach for supplier classification. Manufacturing suppliers can be classified based on multiple factors such as their primary and secondary services, their geographic location, their focus target

market and industry, the products they produce, and the technologies they use. One useful classification method is a process-based classification that creates supplier families based on the similarities in the manufacturing processes they provide. For example, if the providers of Additive Manufacturing (AM) services are identified and grouped together as a family, more targeted and specialized search can be conducted over this subset of the search space. One may choose to further break down the class of suppliers into more specific sub-categories such as providers of metallic and non-metallic AM services. Building more specific and multi-dimensional classes of suppliers improve the precision of the search process. The competing classification methods are first tested through building simple and non-overlapping supplier classes, namely, *casting* service providers and *CNC machining* service providers. The second phase of evaluation involves more similar and complex classes which makes the classification process more difficult.

The term-based text classifier receives the textual narrative of an unlabeled supplier as the input and classifies the supplier under one of these two classes. Casting and CNC machining have different dictionary of technical terms. For human experts, it is readily possible to distinguish machining text from casting text. However, a machine agent needs to be trained before it can make this distinction. Table 3 lists some of the unique technical terms for each category.

Table 3. Example terms related to casting and machining processes

Casting	CNC Machining
Cast, gray, heat, treating, investment, mold, sand, shell, ductile, casting, housing, foundry, iron, molding, sand , core, furnace, drag, die, green sand casting, plate, pattern, ductile, melting, heat treatment	CNC milling, grinding, vertical milling machine, Swiss turning, turret lathe, live tooling, shoulder facing, countersinking, contour cutting, hard turning, precision machining, threading, tapping, straight turning, demurring

Figure 4 shows the overall flow of the term-based classification process. The classifier is built with R-language which has a rich set of libraries and packages for text mining. R is an open source programming language and is widely used for data mining and analysis. Further details related to preparation of training data and classification of test data is provided in the following sections.

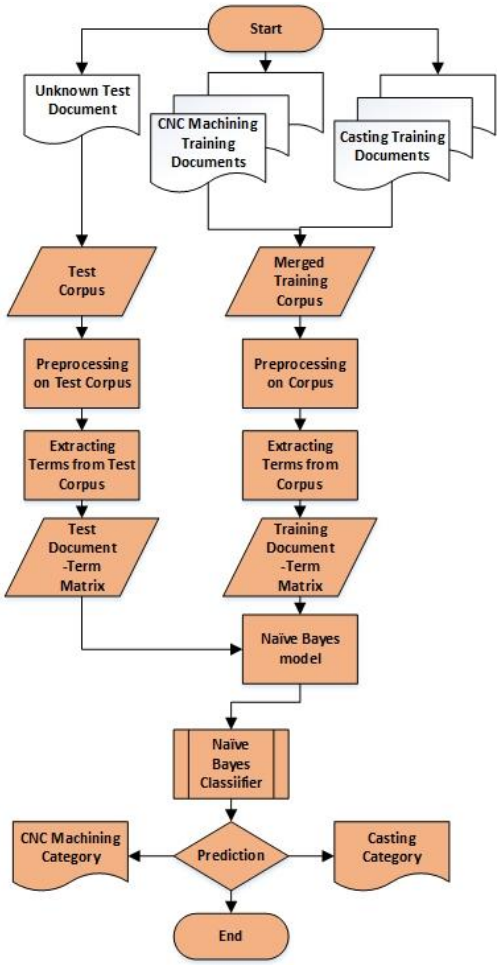


Figure 4. Flowchart of the supplier classifier

Preparation of Training Data

The term-based text classification uses a supervised learning method. In supervised learning, a set of pre-classified, or pre-labeled, text is used as the training data. The training dataset is collected from Thomas Net portal narratives. To build the training dataset, an equal number (m=20) of suppliers from both categories (i.e., casting and CNC machining) are selected and their online profiles are converted into textual documents. These documents collectively form the training corpus. After going through a series of pre-processing steps, as described below, the corpus documents are converted into term

vector representations. The resultant term vector serves as the *bag of terms* or *dictionary* to be used by the classifier.

Pre-Processing

The cleanup process, or *preprocessing* of the corpus, is composed of multiple steps such as removal of English stop words such as “and”, “is”, “for”, “the”, removal of the generic words such as “product”, “capable”, “include”, and “market”, and removal of punctuations, numbers, and whitespaces. Also, stemming is performed to reduce derived terms to their word stem, or root form. The next step in pre-processing is converting compound phrases into single terms. A Naïve Bayes classifier works based on single terms and it is not capable of accommodating compound phrases. However, decomposition of compound phrases will result is loss of meaning. For example, if “Green Sand Casting” is broken down into its constituting atomic terms (i.e., “green”, “sand”, and “casting”), its underlying semantics is lost. To overcome this flaw, the important compound phrases related to casting and machining available in the corpus were replaced by their abbreviations. In this way, they can be treated as single terms. For example, “Electrical Discharge Machining” is replaced with EDM during pre-processing. Figure 5 and Figure 6 illustrate the partial view of the corpus before and after the preprocessing.

Figure 5. Training corpus before preprocessing

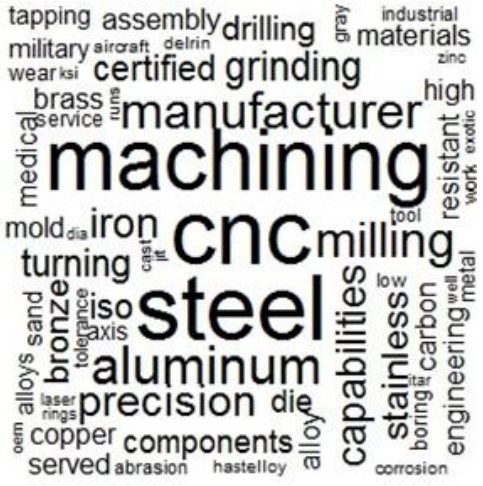


Figure 6. Training corpus after preprocessing

The font size of the terms indicates the degree of importance of the terms in the corpus based on their frequency of occurrence. Through comparing Figure 5 and Figure 6, one can see how key terms with manufacturing significance are gaining more weight in the corpus after preprocessing. The preprocessed corpus has two different classes, namely, $c_{Casting}$ and $c_{CNCmachining}$, each with equal number of training documents (m) as defined below. Let *d* denote a document in class *c*.

$$C = [c_{Casting}, c_{CNC\ machining}] \quad (1)$$

$$c_{Casting} = [d_{Casting_1}, d_{Casting_2}, \dots, d_{Casting_m}] \quad (2)$$

$$c_{CNC} = [d_{CNC\ machining_1}, d_{CNC\ machining_2} \dots d_{CNC\ machining_m}] \quad (3)$$

Creating the Document-Term Matrix

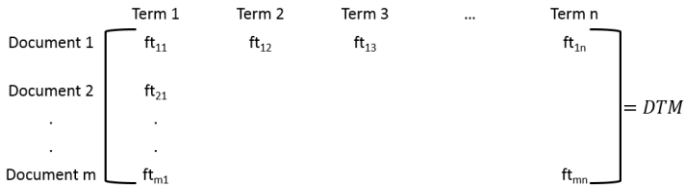
The next step after preprocessing of corpus documents is *feature selection*. It involves selecting the features (terms) in each training document that characterize the classes of interest. Term frequency is used as the criterion for feature selection. Accordingly, the terms that occur more frequently in a pre-processed corpus are deemed more relevant.

These terms form casting and CNC machining *dictionaries*. A partial view of the CNC machining dictionaries based on two types of corpora, namely, ThomasNet and manufacturers' website narratives, is shown in Table 4. As can be seen in this table, ThomasNet term dictionary contains more relevant terms compared to manufacturer website dictionary. For this reason, ThomasNet dictionary was used for this experimentation for building the corpus of the term-based method.

Table 4. Top 20 terms in CNC machining dictionaries

Top 20 terms based on Thomas Net corpus	**Top 20 terms based on manufacturers website narratives**
aerospace	cnc
aluminum	components
assembly	customers
axis	engineering
brass	equipment
capabilities	grinding
certified	high
cnc	horizontal
drilling	industries
grinding	facility
horizontal	machine
iso	machines
machined	machining
machining	manufacturing
medical	milling
milling	process
precision	precision
Stainless	production
Steel	products
turning	vertical
vertical	tool

Once the terms within the dictionary are identified, the Document-Term Matrix (DTM), also known as space vector model, can be built as follow.



The rows of the matrix represent training documents and the columns represent the terms in training documents. The body of matrix presents the frequency of the terms in each document. Terms in the dictionary might be encountered multiple times, some more frequently than others. The vector model of a document represents the document as a set of terms coming from the dictionary of terms.

Test Data

The test data is the portal narratives of the suppliers that are supposed to be classified as either CNC machining supplier or casting supplier. The test data is also collected from Thomas Net. Similar to the training data, the test data needs to be preprocessed and converted into term vector representation. The only difference between test and training documents is that the training documents are permanent data while test documents just go through the classification process and then vanish.

Naïve Bayes Classifier

The proposed method for supplier classification uses Naïve Bayes classifier for document classification. Simplicity, scalability, fast performance are among the advantages of Naïve Bayes. The most important feature of Naïve Bayes classifier is that it needs fewer training data compared to the other supervised techniques (Ting et al., 2011). Naïve Bayes in machine learning is a probabilistic classifier. Naïve Bayes classifier functions based on Bayes rule with independence assumptions. Bayes rule relates the probability of terms used in test documents to the probability of the used terms

in training documents. The independence assumptions considered in this work are as follows:

- The probability that a term **t** appears in a document **d** does not depend on the context of document, it only depends on the class **c** of the document.
- Knowing the previous term t_{k_i} in the document (or any other term) does not alter the probability that a term occurs in position b_i in the document. This is shown in Eqn. (4).

$$P\left(b_i = t_{k_i} \middle| b_j = t_{k_j}, c\right) = P\left(b_i = t_{k_i} \middle| c\right) \quad or \quad P\left(b_i \middle| b_j, c\right) = P(b_i|c) \qquad (4)$$

As mentioned earlier, a document is converted into a sequence of n terms. This can be shown as $d = (t1, t2, \dots, tn)$. For instance, having a simple, one-sentence document as "We provide customized machining, milling grinding and drilling" , d after preprocessing can be written as d=(machining, milling, grinding, drilling). We are interested in finding the probability of a given document belonging to a given class c. This probability is denoted by $P\ (c|d)$. By applying Bayes rule and independence assumption, this probability can be formulated using Eqn. (5).

$$P(c|d) \propto P(c) \prod_{i=1}^{|d|} P(t_i|c),\ P(c) = P\ (t_1, t_2, \dots t_n|c) \qquad (5)$$

In this equation, $P(t_i|c)$ represents the conditional probability of term t_i occurring in a document of class c. In other words, it can be said that $P(t_i|c)$ is a measure of how much proof t_i supplements that c is the correct class. P(c) presents the prior probability of a document which is occurring in class c. If a document's terms do not deliver clear indication for one class against another class, the class which has the higher prior probability will be selected. In text classification, the best class must be assigned to the

document. The Naïve Bayes classifier describes the best class as the most likely or *Maximum A Posteriori* class (MAP) C_{MAP}. This can be calculated by Eqn. (6).

$$C_{MAP} = argmax_c P(c|d) = argmax_c \prod_{i=1}^{|d|} p(t_i|c)\ p(c) \quad (6)$$

However, $P(t_i|c)\ and\ P(c)$ need to be estimated from training documents. They can be estimated based on the Maximum Likelihood Estimate (MLE). Equations (7) and (8) are used for these estimations.

$$p(c) = \frac{Nc}{N} \quad (7)$$

Where N_c denotes the number of document of class c while N is the number of all documents.

$$p(t_i = w|c) = n(w,c)/\sum_{w \in W} n(w,c) \quad (8)$$

To calculate $n(w,c)$, Eqn. (9) is used which declares that word w occurs $n(d,w)$ times in document d.

$$n(w,c) = \sum_{d \in c} n(d,w) \quad (9)$$

Equation (8) estimates the probabilities from the term frequencies in training documents. However, the test document might contain new terms which are not included in training documents or vice versa. For those new terms, the probability of 0 is assigned that makes the computation inaccurate. To overcome this situation, the probabilities were smoothed by applying Laplace correction. Laplace correction adds one to the nominator and denominator of the Eqn. (8) resulting in the following equation.

$$P(t_i = w|c) = (n(w,c) + 1)/\sum_{w \in W}(n(w,c) + 1) \quad (10)$$

This equation is based on the assumption that each word occurs at least once in a document. Based on Naïve Bayes equations described above, given a test document, the

developed tool can determine the corresponding class of the test document probabilistically.

Validation of the Term-Based Classification Method

To evaluate the performance of classifiers, three measures are typically used. These measures are defined as:

- Precision: Out of all documents classified under class c, what portion correctly belongs to class c.
- Recall: Out of all class c documents in the test dataset, what portion is correctly classified under class c.
- F-measure: a compound score which is calculated based on precision and recall and is used to measure the model accuracy.

The above mentioned measures can be calculated using the following equations:

$$Precision = \frac{a}{a+b} \quad (11)$$

$$Recall = \frac{a}{a+c} \quad (12)$$

$$F - measure = 2 * \frac{Precision * recall}{Precision + recall} \quad (12)$$

According to the Table 5, *a* is the number of casting documents that are correctly classified as casting (True Positive) and *b* is the number of non-casting documents that are incorrectly classified as casting document (False Positive). *c* and *d* denote False Negative and True Negative values respectively for casting class. A similar table can be built for CNC machining class.

Table 5. Result table for validation of Casting class

	Predicted Casting	**Predicted Not Casting**	**Sum**
Is Casting	a (True Positive)	c (False Negative)	a+c
Is not Casting	b (False Positive)	d (True Negative)	b+d
			n

The performance of the developed classifier was tested experimentally for different combinations of test and training documents. Figure 7 show the results based on the *recall* measure. The vertical axis represents the number of training and test documents for each class. For example, (20, 15) indicates that there are 20 training documents and 15 test documents are used for each class. As can be seen in the graph, the overall performance of the classifier with respect to recall measure improves with increase in the number of training documents.

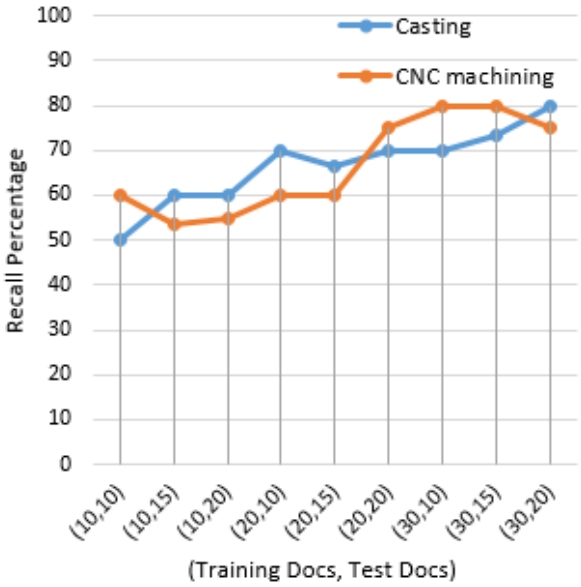


Figure 7. Recall comparison chart

The classifier was also tested for different sets of training and test data based on the F-measure. Table 6 represents the averaged F-measures obtained for different numbers of training and test documents.

Table 6. F-measure results for Standard method

Naïve Bayes Classifier		**Number of Training documents for each process**		
	F-measure	**10**	**20**	**30**
Number of Test documents for each process	**10**	0.54	0.64	0.75
	15	0.56	0.63	0.74
	20	0.57	0.72	0.77

The obtained results confirms that the accuracy of the model increases with increase in the number of training documents. For instance, when the number of training document increases from 20 to 30, the accuracy is improved by %11. The term-based

method proved to be efficient when used for building simple and disjoint classes of manufacturing service providers. However, the following deficiencies were identified through this study:

1) Sensitivity to the size and quality of training data: The accuracy of the classifier tool may decline by adding more training documents to the system. More training documents means a larger dictionary of terms. When dealing with a large number of terms, even after preprocessing, some of the irrelevant terms may remain in the dictionary, thus reducing the overall accuracy of the tool. This implies that different training documents, based on their size and content, may influence the quality of the final results differently. The sensitivity of the classifier to the quality of the training documents is considered to be drawback of the text-based classification technique. This drawback becomes a serious flaw when documents are supposed to be classified under more complex classes that require highly specialized training data
2) Semantic degradation and alteration: When a documents is reduced to a set of atomic terms, the connections of the terms with the conceptual and semantic model of the document is not maintained. Therefore, a vector model of a document is not semantically equivalent to the original document.

3) Cost of training data preparation: Preparation of high-quality training data entails going through numerous supplier websites and choosing the relevant narratives pertaining to the class of interest. This activity becomes more tedious and time-consuming when the objective is to form complex, multidimensional classes that require highly specialized training documents that are not easy to find. Once the suitable training documents are selected, they need to be pre-processed in order to eliminate the irrelevant terms. Even after pre-processing, the training data is not completely clean and it still contains some irrelevant or generic terms which will reduce the precision of classification process.

In the next section, a concept-based method for document classification is introduced that resolves the abovementioned issues through providing a robust and high-quality dictionary of manufacturing terms.

Classification Using Concept-Based Method

In order to overcome the deficiencies of the term-based method, a modified classification method is proposed in this section that uses a formal, hand-made thesaurus for generation of training dictionary. In this method, the *bag of terms* is replaced by the *bag of concepts*. The formal thesaurus that is used in this work is called ManuTerms (Ameri, Urbanovsky & McArthur, 2016), and is based on SKOS (Simple Knowledge Organization System) syntax and semantics. SKOS is a standard model designed for representation of thesauri, classification, taxonomies, or any other form of structured vocabulary ("SKOS Simple Knowledge Organization System Reference", 2016). SKOS is built upon RDF and RDFS and is part of the Semantic Web technology suite. It provides a concept-based view of the vocabulary, in which objects are abstract notions

(i.e., concepts) labeled by terms. SKOS concepts are organized in taxonomies and also can be linked together by non-hierarchical relationships. Concepts in a SKOS thesaurus are categorized under multiple concept schemes. The core components of SKOS include concepts, labels, definitions and semantic relations. Each concept in SKOS has one preferred label and one or more alternative labels. SKOS offers three types of semantic relationships including hierarchical, associative and mapping relationship. Hierarchical relationships are of narrower (more specific) or broader (more general) types. Concepts in SKOS can be related to any other concept within the same thesaurus or mapped to other similar or exact matching concepts in other data sets. Figure 8 shows the concept diagram of the *Molding Sand* concept based on SKOS representations. As can be seen in this graph, *Foundry Sand* and *Casting Sand* are the alternative terms for *Molding Sand*.

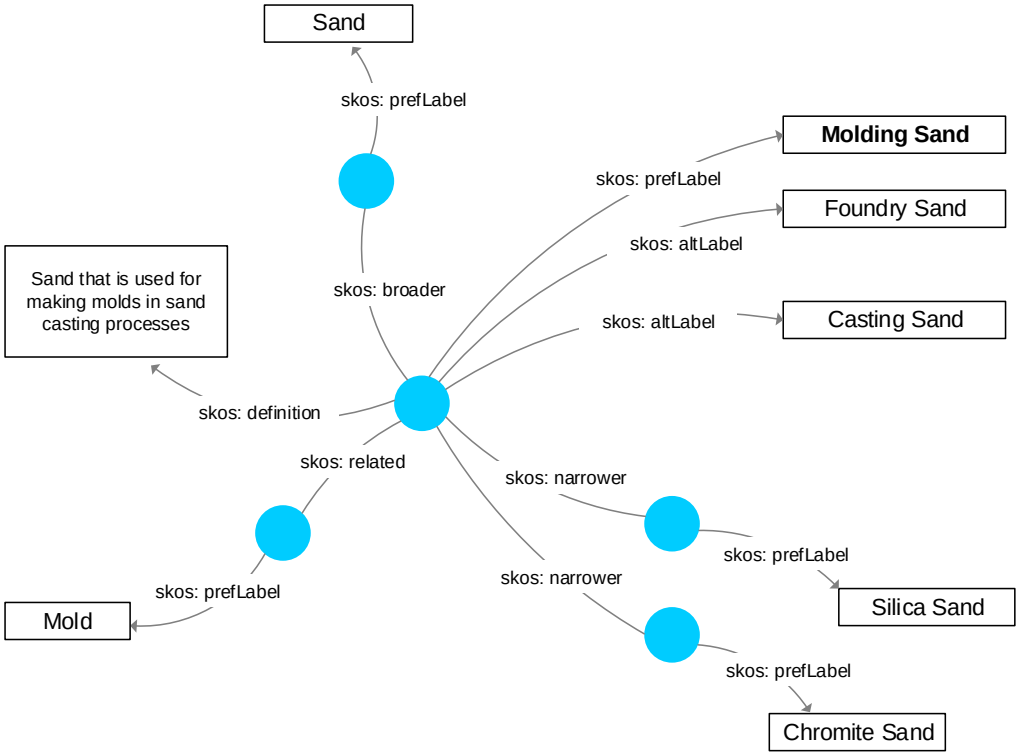


Figure 8. The concept diagram of the molding sand based on SKOS terminology

Figure 9 shows a SKOS excerpt related to the *Mechanical Subtraction* concept in ManuTerms. It provides a definition for *Mechanical Subtraction* in natural language, its alterative and preferable labels and also its narrower or broader concepts. This concept is directly imported from DBpedia, a crowd-source structure information model extracted form Wikipedia. One of the main advantages of SKOS models is that they can be extended and enriched semantically in a decentralized fashion through connecting with different datasets on Linked Open Data [3]. This will significantly reduce the cost of maintaining and extending the reference thesaurus.

[3] http://linkeddata.org/

dbpedia:Machining a skos:Concept ;

skos:prefLabel "**Mechanical Subtraction**"@en ;

skos:narrowerTransitive dbpedia:Arbor_milling , dbpedia:Spot_facing , dbpedia:Ornamental_turning , dbpedia:Laser_drilling , dbpedia:Diamond_turning , dbpedia:Gang_milling , dbpedia:Abrasive_jet_machining , dbpedia:Pencil_milling , dbpedia:Friction_drilling

skos:broaderTransitive skos:prefLabel "**Conventional machining**"@en ;

skos:definition "Conventional machining is a collection of material-working processes in which power-driven machine tools, such as saws, lathes, milling machines, and drill presses, are used with a sharp cutting tool to mechanically cut the material to achieve the desired geometry.

skos:altLabel "Machining"@en , "Subtractive manufacturing"@en , "Mechanical cutting"@en , "Cutting and machining"@en , "Machining of Castings"@en , "Conventional machining"@en , "Subtractive process"@en , "Material removal"@en , "Cutting"@en , "industrial machining"@en , "conventional turning"@en , "mechanical machining"@en ;

dcterms:source dbpedia:Machining ;

dcterms:creator "amerif" .

Figure 9. SKOS source-code excerpt related to Mechanical Subtraction

ManuTerms

ManuTerms is a SKOS-based thesaurus for manufacturing terms. ManuTerms was originally designed for capturing the terminology used by manufacturing companies for describing their technological capabilities. It can be considered as lightweight ontology for manufacturing service description due to its simple semantics. Like any other thesaurus, ManuTerms is a living entity and constantly evolves (Ameri et al., 2014). At

the time of preparation of this text, ManuTerms contained more than 2000 concepts organized under nineteen concept schemes as illustrated in Figure 10.

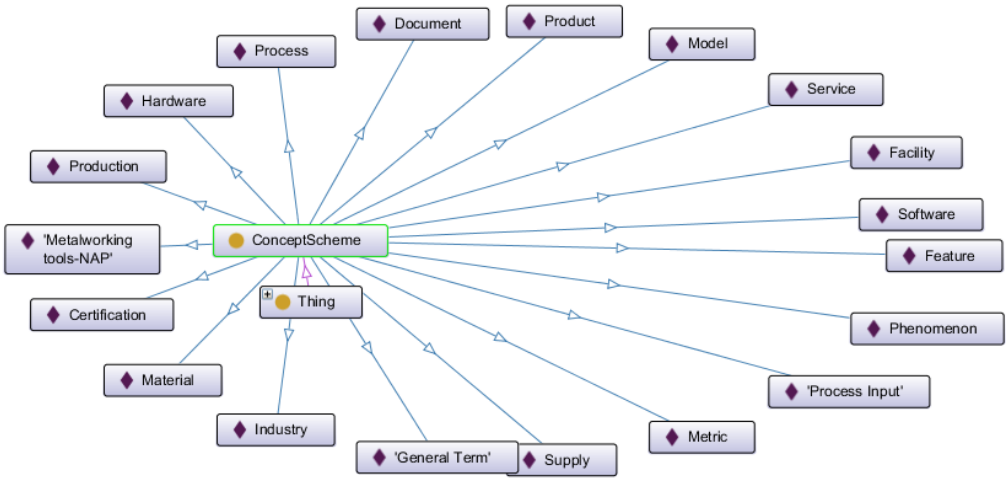


Figure 10. Concept Schemes in ManuTerm

ManuTerms can be automatically updated and extended through the datasets available on linked open data such as DBpedia and GeoNames[4]. Also, it can be continually extended, validated, and enriched by domain experts that contribute to development of ManuTerms core concepts. Therefore, it can be used as a reliable resource for semantic annotation and characterization of a wide range of manufacturing services. In this section, a thesaurus-based service classification method is introduced that extracts its training data from ManuTerms.

Classification Method Based on SKOS

In the concept-based method, the training data is provided through sending queries to ManuTerms. Based on this method, document classification is performed at the conceptual-level, thus enhancing the semantic relevance of the results. Figure 11 shows how ManuTerms is integrated with the proposed classification system.

[4] http://www.geonames.org/

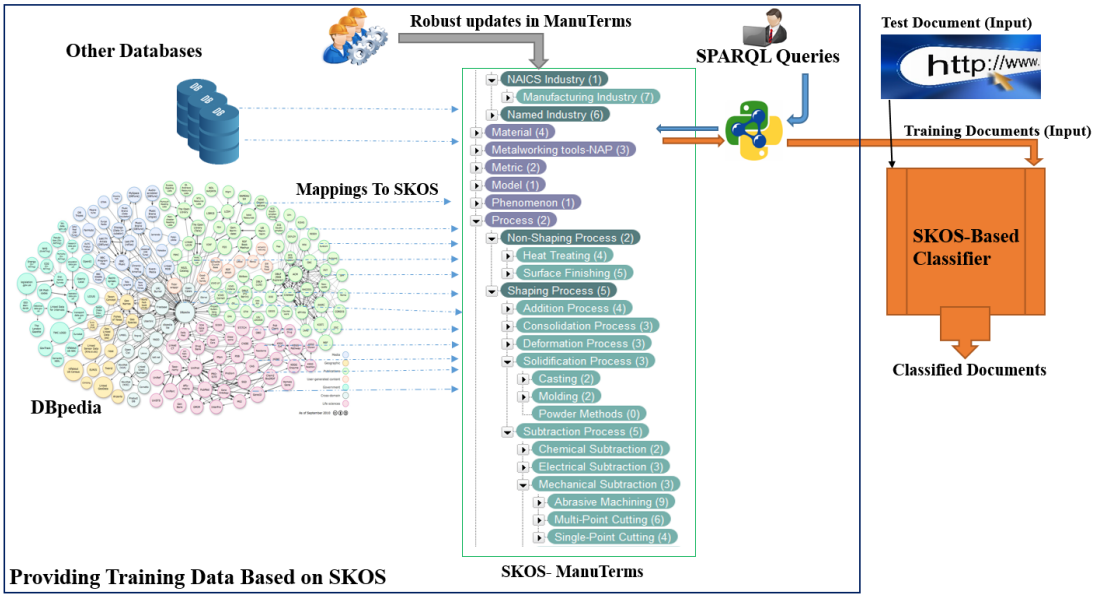


Figure 11. Proposed classifier based on SKOS

Depending on the features of the class of interest, a subset of ManuTerms concepts will be used as the training data for each classification task. The terms that build the training data are the labels (preferred or alternative) of ManuTerms concepts. Therefore, training data maintains its connection with the ManuTerms' semantic graph.

Training Data Preparation

The first step in training data preparation is selection of the relevant concepts for a given class. The selected concepts are the features that characterize a class. This is conducted in two steps.

Step 1: Entry concept selection: The *entry concepts* are those concepts that are used as the entry point to the thesaurus. They are the first set of predicative features for the class. Selection of the entry concepts for a given classification scenario depends on the degree of complexity of the class. For simple classes, such as *machining* class, only a few entry concepts, such as *machining*, *milling*, and *turning* may be used. For more

complex and higher-dimensional classes, more entry concepts are required to fully characterize a class through addressing its different facets.

Step 2: Sub-tree extraction: In this steps, the concepts that are semantically connected to the entry concepts are extracted through submitting multiple SPARQL queries. The semantic connection with the entry concepts is realized through hierarchical or associative relations. The size of the extracted sub-tree depends on the expected levels of precision and recall measures. For high recall and low precision results, only the immediate broader and narrower concepts of the entry concepts are often adequate. For higher precision, deeper subsets of the thesaurus should be extracted through retrieving all related concepts and their narrower and broader concepts in multiple levels. The dictionary of terms is built through aggregating the returned concepts from different SPARQL queries. No pre-processing is required in the concept-based method because the corpus only contains relevant terms with manufacturing significance.

An example SPARQL query and the returned concepts for the query are shown in Table 7.

Table 7. SPARQL query and its partial results

SPARQL Query	Returned
PREFIX skos:<http://www.w3.org/2004/02/skos/core#> SELECT ?label ?x WHERE { <http://dbpedia.org/resource/Machining> skos:narrowerTransitive ?x . ?x skos:prefLabel ?label. }	"Arbor milling", "Spot facing", "Ornamental turning ","Laser drilling","Diamond turning" ,"Gang milling", "Jet Machining", "Pencil milling" ,"Friction drilling", "Grinding","Horizontal Milling","Manual Milling", "Manual Turning", "Swiss Machining" ,"Cut-off", "Honing", "Vertical Milling", "Facing" ,"Grooving", "Lapping"

This query returns the *preferred labels* of the *narrower* concepts of the *Machining*. Another query can be added that extracts the *alternative labels* of all returned concepts from the previous query. Once the terms are selected, the compound terms are converted into a single string of characters since the classifier can only work with single terms. For example, the compound term "Friction Drilling" is first converted into "FrictionDrilling" and then it is added to the dictionary.

The Machining class is considered to be simple classes as it only deals with a single dimension of supplier capability model (i.e., manufacturing process). However, often it is of interest to categorize suppliers under more complex and multidimensional classes. For instance, one might be interested in forming a class of suppliers "*who work for medical industry and have expertise in deep hole drilling and precision machining of small parts which can also make complex assembles*". This class is characterized through multiple predictive features such as *medical industry*, *deep hole drilling*, *precision machining*, *complex assemblies*, and *small parts*. Building complex classes requires more specialized dictionary of terms. In the term-based method, the dictionary of terms is harvested, through machine learning, from pre-classified supplier narratives. The challenge is that building specialized dictionaries of terms in the term-based method entails finding highly specialized capability narratives which could be a tedious task. This will result in having very few positive training examples for the class. However, ManuTerms already contains multiple taxonomies of specialized manufacturing terms. The information-content, or specificity, of terms increases with increase in the hierarchical depth of terms. Therefore, a specialized dictionary of terms that accurately characterizes a complex class can be built through selecting a subset of ManuTerms that point to various facets of the class of

interest. This selection is done through submitting suitable SPARQL queries against ManuTerms. Each query pertains to one of the features (i.e, entry concepts) of the complex class. Table 8 shows some example SPARQL queries and their results for the complex class example provided above. A union of the returned terms by different queries forms the training data for the class of interest. The entry concepts used in this example include, *Complex Assemblies*, *Medical Equipment* and *Supplies Manufacturing*, *Deep Hole Drilling*, *Small Parts*, and *Precision Machining*.

Table 8. Sample queries for complex classes

SPARQL Query	Description	Result

PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
SELECT ?x
WHERE {
<http://infoneer.poolparty.biz/Processes/391> skos:altLabel ?x .
}
	A simple query which returns few terms related to *complex assemblies.*	"complex geometries"
		"complex geometry"
		"complex parts"
		"Complex Machined Components"

PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
SELECT ?label ?x
WHERE {
<http://infoneer.poolparty.biz/Industry/9> skos:narrower ?x .
?x skos:prefLabel ?label.
}
	It returns narrower terms related to the medical equipment and supplies manufacturing.	"Surgical and Medical Instrument Manufacturing"
		"Surgical Appliance and Supplies Manufacturing"
		"Dental Equipment and Supplies Manufacturing"
		"Ophthalmic Goods Manufacturing"
		"Dental Laboratories"

PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
SELECT ?label ?m
WHERE {
<http://infoneer.poolparty.biz/Processes/238> skos:broader ?x .
?x skos:narrower ?m.
?m skos:prefLabel ?label .
}
	It returns broader terms for *deep hole drilling*.	"Pilot hole" "Hole punch"
		"Boring" "Single-pass bore finishing" "Drilling"
		"Drill bit shank" "Orbital Drilling"
		"Tapping"
		"Counterboring"
		"Countersinking"
		"Reaming"
		"Deep-hole drilling"

Concept Weighting

Since each term, single or compound, appears in the thesaurus only once, frequency cannot be used as before as the metric that indicates the importance of the terms in the context of the class of interest. Instead, a customized weighting scheme, as shown in Figure 12, is used for specifying the importance of each term in the dictionary. For example, the weight of the preferred label of the entry concept is 9 whereas, the alternative labels receive a weight of 5. The preferred and alternative labels of the narrow, broader, and related concepts within ManuTerms are weighted 3 and 1 respectively. A weight of 9 is also assigned to the exact match of the entry concept in an external thesaurus on the Linked Open Data.

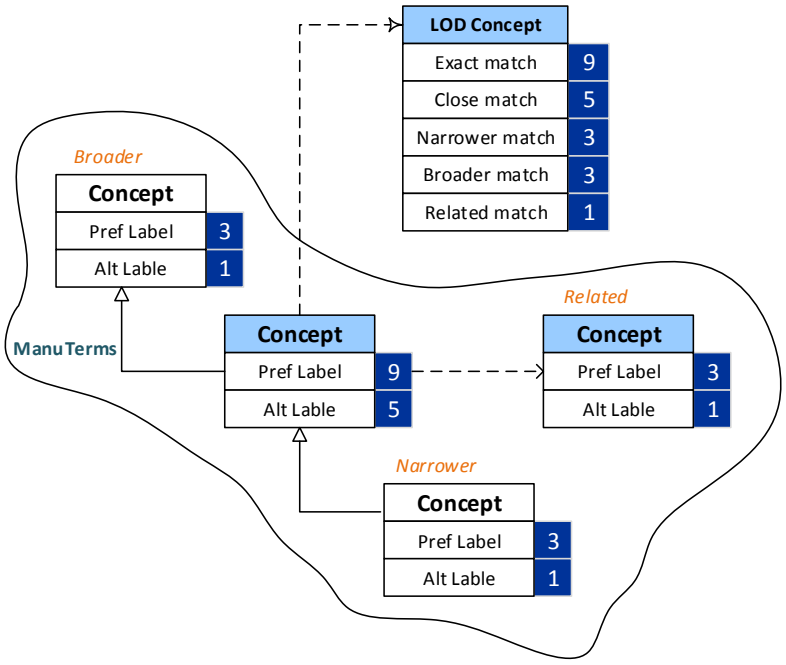


Figure 12. The term weighting scheme used in the concept-based method

Test Data Preparation

Similar to the text-based method, test data is originally represented as natural language text extracted from service providers' websites. The text is then pre-processed

and converted into the vector model. During pre-processing, the compound terms are converted into single terms such that they can be used by the classifier. This conversion is applied only on the compound terms that can be found in the thesaurus.

Comparison Between Term-Based and Concept-Based Methods

In order to compare the performance of the term-based and the concept-based classification methods, an experiment was designed and conducted based on two different datasets. As can be seen in Table 9, the first dataset was composed of 42 documents equally divided between milling and turning classes (similar classes) and the second dataset was comprised of 42 documents belonging to casting and machining classes (dissimilar classes). The expectation was that the two methods should have comparable performances when classifying documents under two disjoint classes. However, as the classes become more similar, accurate classification becomes more challenging for classifiers. The hypothesis to be tested in this experiment was that the concept-based classifier performs better when classifying documents under relatively similar classes with similar vocabularies.

Table 9. The experimental data sets

Data Set	Source	Classes	Number of Docs in each classes	Total Docs in Data set	Total Words in Data set
Similar data set	Websites of suppliers	Milling, Turning	21	42	2090
Dissimilar data set	Websites of suppliers	Casting, CNC machining	21	42	1805

Results for Term-Based Method

Under each scenario, 4 random documents from each class were selected to be used as training data and the remaining 38 documents were treated test data. To avoid any bias caused by the particular choice of training data, the experiment was run in 5 iterations. In each iteration, 4 random documents of each classes were selected as the training data and subsequently, the prediction was performed on the remaining documents. Table 10 and Table 11 summarize the results of for the term-based method for dissimilar and similar class scenarios with respect to precision, recall, and F-measure.

Table 10. The summary of results for dissimilar classes based on the term-based method

Iteration	Recall	Precision	F-Measure	Accuracy
1	0.88	0.83	0.85	0.85
2	0.94	0.88	0.90	0.91
3	0.94	0.80	0.86	0.85
4	0.82	0.93	0.87	0.88
5	0.88	0.88	0.88	0.88
Mean	0.89	0.86	**0.87**	0.87

Table 11. The summary of results for similar classes based on the term-based method

Iteration	Recall	Precision	F-Measure	Accuracy
1	0.76	0.65	0.70	0.67
2	0.58	0.66	0.61	0.64
3	0.64	0.78	0.70	0.73
4	0.52	0.56	0.54	0.55
5	0.70	0.60	0.64	0.61
Mean	0.64	0.65	**0.63**	0.64

Different F-measure values obtained for different iterations is an indication of the sensitivity of the term-based method to the choice of training data. This variation is more significant in the similar class scenario.

Results for Concept-Based Method

The training data for the concept-based method was obtained through sending multiple SPARQL queries to ManuTerms. Only one iteration per scenario was needed for the concept-based method because the SPARQL queries always return the same set of concepts to be used as the training data. Table 12 shows the entry concepts and SPARQL queries used for building the training data. Because the classes used in this experiment are considered to be simple classes, only a few entry concepts were needed to adequately characterize the class and retrieve the relevant concepts.

Table 12. The entry concepts and SPARQL queries used for building training data

Class	Entry Concepts	Query types	# of returned concepts
CNC machining	CNC Machining Abrasive Machining Machining Service	Broader(prefLabel),Narrower (prefLabel), Related (prefLabel), Narrower-of-narrower (prefLabel), Narrower(altLabel)	98
Casting	Casting Ceramic Mold Casting Expendable Mold Casting	Broader(prefLabel),Narrower (prefLabel), Related (prefLabel), Narrower-of-narrower (prefLabel), Narrower(altLabel)	49
Milling	Milling	Broader(prefLabel),Narrower (prefLabel), Related (prefLabel), Narrower-of-narrower (prefLabel), Narrower(altLabel)	67
Turning	Turning	Broader(prefLabel),Narrower (prefLabel), Related (prefLabel), Narrower-of-narrower (prefLabel), Narrower(altLabel)	43

The results of the Concept-Based Method for both similar class and dissimilar class scenarios are shown in Table 13.

Table 13. The summary of results based on the concept-based method

Dissimilar Classes	**Recall**	**Precision**	**F-Measure**	**Accuracy**
	0.93	0.88	**0.90**	0.91
Similar Classes	Recall	Precision	F-Measure	Accuracy
	0.80	0.84	**0.821**	0.825

Discussion

Based on the results presented in Table 11 and Table 13, and as shown in Figure 13, both methods perform equally well under dissimilar class scenario.

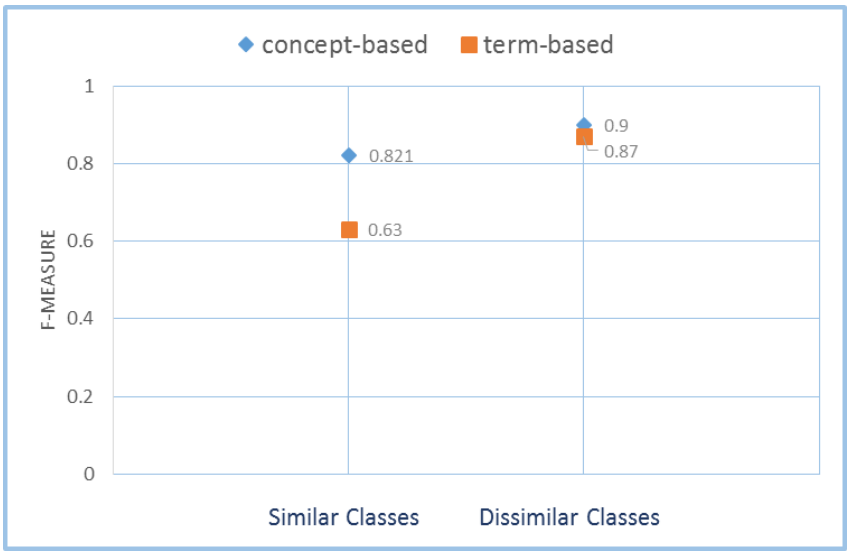


Figure 13. Comparison between the results obtained from concept-based and term-based methods

It is an expected outcome because simple dissimilar classes of Casting and CNC machining have distinct vocabularies. Therefore, most classifying methods can assign a label to an incoming document with a fairly high confidence level. However, it should be noted that the concept-based method obtained slightly better results using a smaller dictionary of terms compared to term-based method. The concept-based method proved to be more accurate when classifying similar documents with overlapping vocabularies.

The improved performance of the concept-based method can be attributed to using a dictionary of terms that are semantically and contextually related to the key features of their corresponding classes. The term-based method breaks the compound phrases into atomic terms, which results information loss. This also results in an undesirable growth in the size of the dictionary of terms that would increase the computational costs in large-scale applications. This issue is resolved in the concept-based method since the training data maintains its semantic relevance by staying at the conceptual level while minimizing the size of the training data. In addition to being more accurate, the concept-based method is less sensitive to the noises induced by random terms in the training data. The SPARQL queries used in this experiment for extracting relevant concepts only return the concepts that are at most two levels deeper than the entry concepts. However, to improve the accuracy, at the expense of recall, one can increase the depth of the queries.

Summary

Manufacturing suppliers are increasingly strengthening their web presence in order to improve their visibility and remain competitive in the global market. With the explosive growth of unstructured content on the web, more advanced methods for information organization and retrieval are needed to improve the intelligence and efficiency of supplier search process. In this chapter, a technique for automated characterization and classification of manufacturing service suppliers based on their textual portfolios is presented. A probabilistic technique that adopts Naïve Bayes method is used as the underlying mathematical model of the proposed text classifier. To improve the semantic relevance of the results, classification is conducted at the conceptual-level rather than at the term-level that is typically used by traditional text classifiers. The necessary steps for

training data preparation and representation related to manufacturing supplier classification problem are delineated. The proposed classifier is capable of forming both simple and complex classes of manufacturing service providers. The performance of the proposed classifier is evaluated experimentally based on the standard metrics such as precision, recall, and F-measure. It is concluded that the proposed concept-based classification technique outperforms the traditional term-based methods with respect to accuracy, robustness, and consequently cost.

III. A HYBRID UNSUPERVISED METHOD FOR MANUFACTURING TEXT MINING BASED ON DOCUMENT CLUSTERING AND TOPIC MODELING TECHNIQUES

Introduction

The abundance of online manufacturing information has resulted in creation and continual expansion of an unstructured and informal manufacturing datasets. Using data analytics, and in particular, exploratory text mining techniques, one can extract interesting and non-trivial knowledge patterns hidden in the data available online. The extracted knowledge can be formalized and imparted to manufacturing ontologies and information models to improve the intelligence of various decision support systems and business solutions. For example, through mining technical blogs, the trending topics in manufacturing can be discovered that might shed light on new emerging technologies, markets, or product features. Also, through building supplier clusters with similar capabilities, new suppliers can be searched in a more targeted manner.

There are multiple text mining techniques, such as summarization, classification, clustering, topic tracking, and association rule mining that can be applied to the manufacturing websites. In the previous chapter, a supervised text classification technique based on Naïve Bayes algorithm was introduced for building classes of manufacturing suppliers based on their technological capabilities described in natural language. The proposed classifier was evaluated experimentally and its accuracy was measured using precision and recall metrics. The developed classifier can significantly improve the performance of online supplier discovery engines. However, preparation of high quality training data for machine learning presented to be a challenge. Another caveat of supervised classification is that it requires a priori assumptions and conceptual

models about the data. However, text mining techniques become significantly more valuable if they can discover the pieces of knowledge and insight that were unpredictable otherwise. To take a step further toward enabling exploratory text analytics in the manufacturing domain, two unsupervised text mining techniques, namely, clustering and topic modeling, are adopted in this work. *Clustering* is the process of grouping documents into clusters based on their content similarity (Manning, Raghavan & Schütze, 2008). *Topic modeling* (Blei, 2012) is a method for finding recurring patterns of co-occurring words in large bodies of texts. The primary objective of the research presented in this chapter is to use text mining techniques for building clusters of manufacturing suppliers and identifying the core concepts that form the underlying theme of each cluster. The secondary objective is to use the extracted terms and concepts from each cluster for extending an existing formal thesaurus of manufacturing concepts.

This chapter is organized as follows. The next section discusses the relevant literature in text clustering and topic modeling. Then the proposed hybrid method for clustering and topic modeling is presented. The later part of this chapter deals with a proof-of-concept experimentation and validation.

Related Work

Text analytics is a relatively new discipline in information technology that has already proven to be efficient in multiple application areas ranging from pharmaceutical drug discovery to spam filtering and summarizing and monitoring customer reviews. In the manufacturing domain, however, it is a relatively new undertaking.

Dong and Liu (2006) proposed a tool for manufacturing website classification and topic modeling in order to reveal hidden themes in their test data. Their proposed website

classifier works based on SVM algorithm. However, SVM is a supervised technique which requires high quality training data. Therefore, in absence of well-prepared training data, the proposed approach will not yield the expected outcome. To resolve this issue, researchers have adopted unsupervised approaches. Using unsupervised techniques, hidden patterns can me revealed in textual documents in absence of any training document. Topic models (Blei, 2012) and Clustering (Manning, Raghavan & Schütze, 2008) are two prominent unsupervised methods for text classification and mining. While Clustering is a long existing technique, Topic Modeling is considered as a relatively new method. Topic models are recently being used to discover and predict the underlying pattern of textual data. Probabilistic Latent Semantic analysis known as PLSA is one of the first topic models which was coined by Hofmann (1999). PLSA is a statistical model which uses the latent variable of topics to relate words to documents (Steyvers & Griffiths, 2005). This model assumes a document is a combination of various topics. Therefore, by having a small set of latent topics or variables, the model can generate the related words of particular topics in a document. One successful application of PLSA is in the bioinformatics context where it is being applied for prediction of Gene Ontology annotations (Masseroli, Chicco & Pinoli, 2012). However, PLSA can suffer from over fitting problems (Alghamdi & Alfalqi, 2015). Latent Dirichlet Allocation (LDA) extends the PLSA generative model. In LDA method, every document is seen as a mixture of different topics. This is similar to the PLSA, except that topic distribution in LDA, has a Dirichlet prior which results in having more practical mixtures of topics in a document (M. Blei, Y. Ng & I. Jordan, 2003). LDA, as a tool for topic modeling, has been used in different applications. For instance, AlSumait, Barbará and Domeniconi (2008) discuss a

LDA-based topic modeling technique that automatically finds the thematic patterns on Reuters dataset. Their proposed approach can be applied in an online fashion as well. However, their approach cannot be directly used for improving the performance of a search engine since their experimental datasets are not the output of a search engine and is a prepared collection. Furthermore, LDA method is applied for public sentiments and opinion mining in product reviews (Shulong Tan et al., 2014; Zhai, Liu, Xu & Jia, 2011). LDA-based topic modeling also is used for exploration of offline historical corpora (Yang, Torget & Mihalcea, 2016; Hu, Boyd-Graber, Satinoff & Smith, 2013).

Most of the existing methods are considered as a standalone applications which are proposing either clustering or topic modeling techniques to help users categorize existing information and infer new information from their own internal corpora. This chapter proposes a hybrid model based on clustering and topic modeling methods to facilitate online search and organization of textual documents and also extraction of thematic patterns in manufacturing corpora.

A Proposed Method for Hybrid Clustering and Topic Modeling in Manufacturing Corpora

The standard method for information search and retrieval is the keyword-based method. In a supplier search scenario, a customer from medical industry who is looking for precision machining services, can simply use precision machining and medical equipment as the search keywords in a generic search engine. Nevertheless, the sheer size of the returned set would reduce the value or information content of the search result. One way to enhance the usefulness of results is to present them to the user as chunks or clusters of documents and then characterize each cluster using a set of features or themes. In the precision machining example, a cluster characterized by features such as precision

machining, medical industry, inspection, and assembly would be of interest for the user if inspection and assembly were the secondary services that the user is looking for. This section proposes a hybrid technique, which facilitates clustering and characterization of the documents available in a large manufacturing corpus. The overall structure of the proposed approach is summarized in Figure 14.

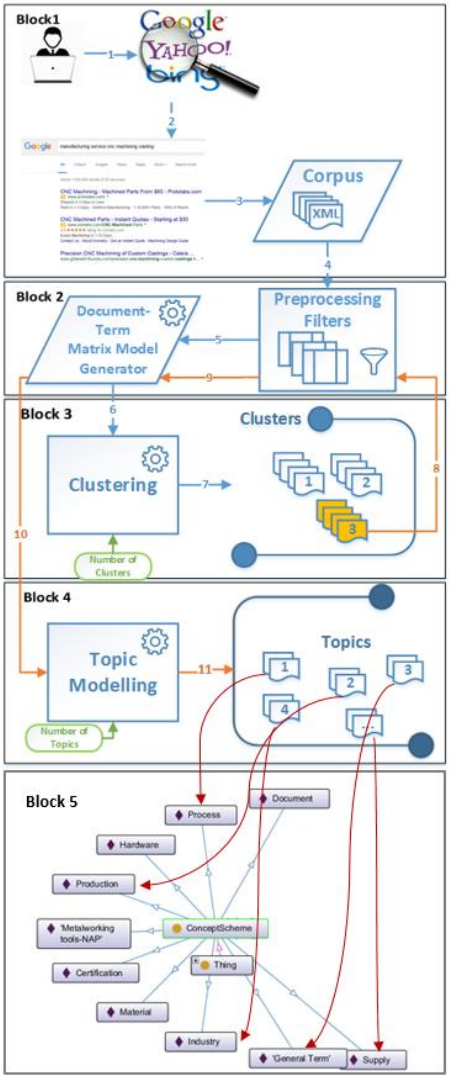


Figure 14. The proposed Hybrid Classifier

This flowchart is composed of 5 Blocks with different functionalities. The implementation of these blocks are discussed in more details in the next sections. K-means in Block 3 and LDA in Block 4 are algorithms that have been serialized in the proposed method. Before discussing the functionality and implementation of each blocks in details, mathematical models of K-means and LDA algorithm need to be briefly introduced.

K-Means algorithm

The k-means algorithm is recognized to be very effective in clustering large textual datasets. K-means first was introduced by MacQueen (1967). The K-Means is an unsupervised algorithm that clusters a set of documents into k clusters based on the vectored attributes of each document. K is a predefined constant determined upfront by user. The objective of this algorithm is to minimize the average squared Euclidean distance of documents from their cluster centers. According to the Eqn. (13) a cluster center is defined as the *centroid* u of the documents in a cluster w.

$$\vec{u}(w) = \frac{1}{w} \sum_{\vec{x} \in w} \vec{x} \qquad (13)$$

RSS (Residual Sum of Squares) is a quantity which determines how well the centroids can represent the participants of their clusters. RSS is defined as "the squared distance of each vector from its centroid summed over all vectors" and can be calculated with Eqn. (14) and (15).

$$RSS_k = \sum_{\vec{x} \in w_k} |\vec{x} - \vec{u}(w_k)|^2 \qquad (14)$$

$$RSS = \sum_{k=1}^{k} RSS_k \qquad (15)$$

w_k is document cluster k, $\vec{u}$ is centroid of the documents in cluster w_k and $\vec{x}$ is document vector in cluster k. it is again worth to mention that document vectors are built based on the DTM discussed in prevoius section.

The objective of K-means algorithm is to minimize RSS through modifying the center of clusters.

LDA Algorithm

Blei et al. (2003) have introduced LDA as a generative probabilistic model. This model can be used for topic modeling of documents in a corpus. The idea behind LDA is that documents of a corpus are characterized as mixtures of latent topics while every topic is characterized by word distribution. LDA is built based on a suite of algorithms that discover thematic structure or topics in documents. Two Basic assumption of LDA are as below.

1. There are only a fixed number of patterns of word in a corpus which are called ***Topics***.

2. Each document of a corpus is made of all topics but with varying probabilities.

The probability of each topic can be inferred from Eqn. (16).

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{Z_{dn}} p(z_{dn}|\theta_d)\, p(w_{dn}|z_{dn},\beta) \right) d\theta_d \qquad (16)$$

Where,

- D is the corpus
- M is number of Documents
- N is number of Words

- θ_d is a document-level variable
- α is a Dirichlet parameter
- β is a Dirichlet parameter
- $z_{dn}\ and\ w_{dn}$ are word-level variables

However, LDA is a very sophisticated algorithm and its detailed description is outside the scope of this thesis work. LDA can help to develop new techniques to search, surf and summarize enormous archives of texts.

Implementation

In order to implement 5 Blocks of proposed technique, different tools and techniques are used. Table 14 summarizes the most important tools and techniques used in this section in more details.

Table 14. Main technologies and tools used for implementation of proposed technique

No.	Tool/Technology	Version	Comment
1	R	0.98.10	Programming Language
2	topicmodels	0.2-2	R Package for Topic Modeling algorithms
3	cluster	2.0.4	R Package for clustering Algorithms
4	e1071	1.6-7	R Package for classification Algorithms
5	TM	0.6	R Package for Text mining
6	XML	3.98	R package for encoding documents
7	WordCloud	2.5	R package for visualizations
8	SnowballC	0.5.01	R package for stemming

Implementation of each blocks are discussed subsequently in the following sections.

Block 1: Building the Corpus

The first step is to create a corpus of manufacturing documents that provides test data in this work. To collect relevant websites, simple web search based on a few keywords is used. In this work, text analysis is carried out on CNC Machining and Casting websites. Therefore, keywords such as machining service, casting service, milling, turning, and sand casting were used for building the corpus. Although this initial web search targets the websites of manufacturing suppliers, typically the result contains technical blogs or articles as well. It is expected that the clustering algorithm can categorize the returned websites based on their content. Each document (i.e., website) in the returned set was converted into a text-only document with Extensible Markup Language format (XML). XML is a markup language which is both human and machine readable. Moreover, an XML file, can contain metadata and due to its generality and simplicity, can be used across different platforms and applications. Figure 15 illustrates an example of a website that is converted to an XML file format.

```
<?xml version="1.0" encoding="UTF-8"?>
<Info>
    <Type>Casting</Type>
    <text> ISO 9001:2008 certified manufacturer
        of castings including machined finished
        castings. Capabilities include precision
        manufacturing, designing, building,
        repairing, milling, lathe work, assembly,
        grinding, metal stamping, EDM, welding,
        turning, reverse engineering, injection
        molding, CAD, custom labeling, pad
        printing silk screening. Kan Ban vendor
        managed inventory programs available.
        On-time delivery. Custom manufacturer
        of castings in alloys including
        continuously cast gray ductile iron, 6061
        T6 aluminum, SAE 660 bronze , chrome
        1045, 5041, 1018 1117 steel. Capabilities
        include finished machining of parts from
        0.5 in. to 8.0 in. dia., centerless grinding,
        boring, rough turning, cut-to-length plate
        cutting . Mid to high-volume production
        capabilities from 100 to 100,000 piece
        runs. Rods, bars, bearings, bushings,
        forgings, plates sheets are also available.
        </text>
</Info>
```

Figure 15. Representation of a document in the corpus

The corpus used in this chapter contains one hundred XML documents that are results of a basic Google search. Table 15 provides more detailed information about the generated corpus.

Table 15. Corpus metadata

Dataset	Source	Number of Docs	Number of unique words
Manufacturing related processes	Supplier's websites	100	7470

Block 2: Customized preprocessing of the corpus

Corpus documents need to be noise-free before they can be further mined. Corpus preprocessing entails removing the redundant and less informative terms in order to create a clean corpus. The first preprocessing step includes removing numbers, punctuation, nonstandard symbols and transforming all words to the lower cases. The

next step is to filter out the stop words and whitespaces. There are two categories of stop words that are removed from the documents during the preprocessing step. The first category belongs to the common words and pronouns such as "I", "we", "the", "and". The other category of stop words is related to the words that frequently appear in manufacturing websites but has marginal information value. Words such as "quote", "inquire", "call", "application", "type", "request", "contact", and "address" belong to this category.

Stemming the words of the documents is the next step. For instance, in a document, terms such as "machinery" and "machining" are stemmed to "machin". This step is necessary for reducing the dimensionality of data and improving the computational efficacy of the text analytics algorithms. **Figure *16*** shows a snapshot of the preprocessed corpus.

Figure 16. A snapshot of a preprocessed corpus

The last preprocessing step is to generate the Document-Term Matrix (DTM) for the manufacturing corpus. DTM is a matrix in which it lists the frequency of the words (or terms) in the manufacturing documents. In the DTM, documents are denoted by rows where the words are represented by columns. If a word repeats in a specific document

for n times, value of its corresponding cell in the matrix is n. This DTM matrix is considered as a vector model which is required by all text mining and machine learning techniques and it is the input to Block 3.

Block 3: Clustering document

This step involves creating groups of similar documents in the corpus. This can be readily done if the corpus is small enough to be surfed manually or a supervised and predefined classification scheme fits the document's collection in the corpus. Then it would be simple to scan the keywords in the documents and classify them according to identified keywords. However, for large corpora, manual classification is a laborious task.

In this work, a K-Means clustering algorithm is implemented which automatically clusters the documents of the corpus in a way that documents in the same cluster are more similar to each other compared to those in other clusters. In K-Means clustering technique, the user needs to specify the number of cluster (K) upfront. Consequently, K document clusters are generated in such a way that the distances of each member of cluster from its own cluster's centroid is minimized. The distances are calculated based on the projection of multidimensional DTM on Euclidean planes. The main steps of the clustering algorithm are once again simplified and listed below.

1. Randomly distributes the documents among the predefined K clusters.
2. Calculates the position of the centroid of each cluster.
3. Calculates the distance between each documents and also each centroids
4. Next, assigns every document to a cluster in which the document is closer to the centroid of that cluster.

5. Steps above are iterated until all documents are assigned to a cluster and no document is relocated to a new cluster.

As mentioned in the previous section the number of clusters (k) has to be specified by users upfront. There are some cases in which the user may have an assumption about the number of clusters in a dataset. Also, it might be the case that the user has absolutely no idea about the content of the corpus and consequently, cannot have an accurate estimate about the number of clusters. Therefore, it is useful to provide the user with an initial estimate about the appropriate number of clusters in a dataset that will result is a meaningful partitioning of the population under study. To estimate the proper number of clusters in the dataset in this work, the Sum of Squared Error (SSE) method is used. SSE refers to the sum of the squared distance between each document (member) of a cluster and its own centroid of the cluster. The corpus holds 100 documents. Therefore, the value of K ranges from 2 to 99. The challenge is to select the proper number of cluster through investigating the SSE corresponding to each cluster. Generally, as it is depicted in Figure 4, when the number of clusters increases from 2 to 99, the SSE is decreased since the clusters become smaller in size. Figure 17 illustrates the SSE plot for this range of clusters.

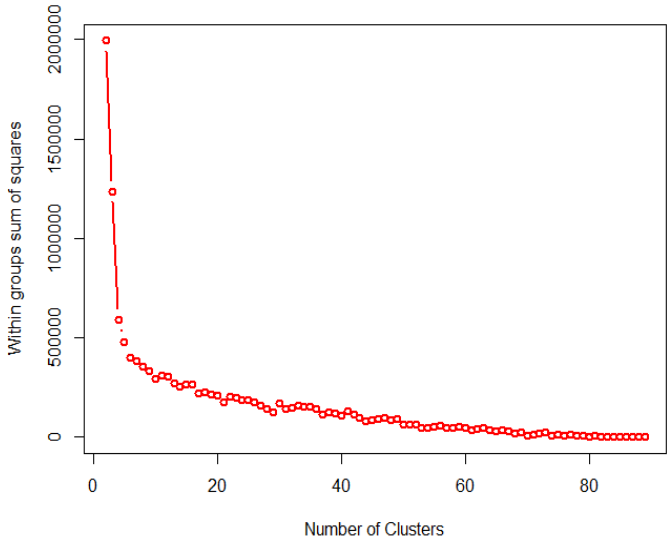


Figure 17. SSE curve for different values of k

Based on the SSE plot, the suggested number of clusters is determined by the elbow point of the SSE curve. As can be seen in Figure 17, the elbow point occurs where the number of clusters is equal to 3. Therefore, three clusters were generated for this particular dataset. These three clusters with their assigned manufacturing documents are illustrated in Figure 18.

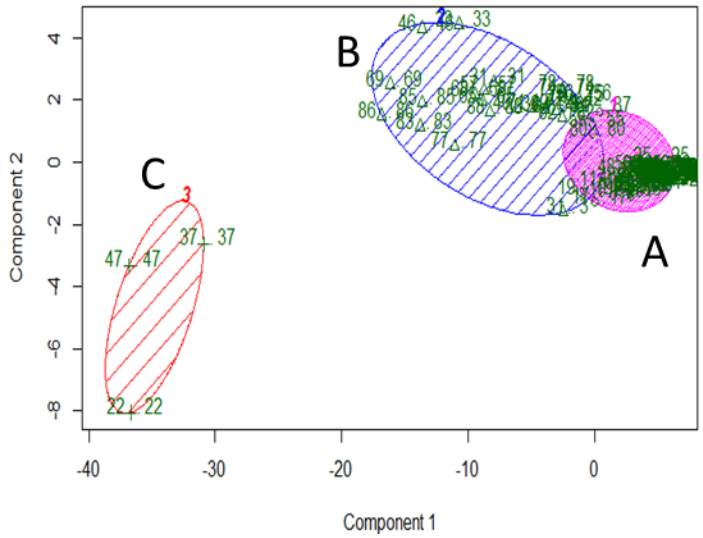


Figure 18. Result of clustering

The plot of clustering result, shown in Figure 18, is obtained based on a dimension reduction technique called Principle Component Analysis (PCA). The proposed clustering algorithm is based on the number of words (or diminutions) in the corpus (7470 words or dimensions) which makes it impossible to visualize the clusters. To overcome this problem, PCA is used to reduce the number of dimensions to 2, thus enabling better visualization.

As it can be seen in the plot, the two upper right clusters (clusters A and B) have partial overlap and are very close to each other while the third cluster (Cluster C) in the lower left corner is clearly distinct from the other two. The members of the overlapped clusters (A and B) are the websites of contract manufacturers who offer machining and casting services. The distinctive feature of the overlapping clusters is the depth of information provided by the member websites. The websites in cluster A contain general and high level information about the type of process and services they offer while the websites in the cluster B provide more in depth information about the type of processes, sub-processes, secondary services, and materials offered by the company. On the other hand, cluster C, mainly includes trade websites, blogs or technical papers. This experiment demonstrated that the clustering algorithm can successfully build meaningful clusters based on the type of documents and also the level detail incorporated in them. However, the clustering algorithm did not make a distinction between machining and casting websites. To further analyze and explore each cluster, topic modeling technique is used in the next step. Cluster B, which contains 50 documents, is selected as the input for topic modeling process.

Block 4: Topic Modeling

As mentioned in the previous section, the input of the Topic Modeling block is a dataset made of 50 documents containing 2313 unique terms. LDA is the algorithm which is implemented in this block and is used for information retrieval and knowledge extraction. LDA technique can be used for automatically discovering abstract topics in documents and classifying each document in the cluster based on the discovered topics. A topic is defined as a group of words that frequently appear together. LDA technique assumes that each document in the dataset is randomly composed of combination of all available topics in the whole dataset. As mentioned in the previous section, LDA is considered as a complex algorithm and can be further studied through Blei et al. (2003). However, the basic steps of the LDA technique are listed below:

1. A desired number of topics, t, is defined upfront.
2. For every manufacturing document, allocate each term or word in that document to one of the t topics.
3. This allocation task initially provides topic representations in all the documents beside distributions of words in all the topics.
4. Next, improve the allocation task by calculating the ratio of words in document d which are already allocated to the topic t, and also by calculating the ratio of words allocations to topic t by considering entire documents.
5. Reassign words to the new topics. The steps above are repeated for multiple iterations to achieve to the optimal result.

As the last stage, the application is run to find two topics in the dataset of 50 documents. Table 16 shows these two topics and their 10 most frequent terms.

Table 16. Top 10 stemmed terms in Topic 1 and Topic 2

	Topic 1	Topic 2
1	cast	turn
2	sand	cnc
3	custom	servic
4	manufactur	part
5	qualiti	industri
6	process	steel
7	mold	tool
8	product	equip
9	die	product
10	aluminum	Precisi

Diverse information can be extracted from this table. For instance, it can be inferred that Topic 1 is mainly about casting processes while Topic 2 corresponds to the machining processes. However, as mentioned earlier, each document can provide information about more than one topic. The proposed method addresses this by returning topic probabilities associated with each document. Table 17 lists these probabilities for first four documents of the cluster B.

Table 17. Documents and their topic probabilities

Document	Topic 1	Topic 2
1.xml	0.384	**0.615**
2.xml	0.126	**0.873**
3.xml	**0.767**	0.232
4.xml	**0.816**	0.183

From Table 17, it can be concluded by the user that the second document, as it is illustrated in Appendix A, belongs to topic 2 which is mainly about CNC machining services. Also, the calculated probabilities suggest that the fourth document belongs to topic 1 which corresponds to the casting process and services.

The performance of the proposed technique can be improved in time through adding more terms to the list of stop words that will be filtered out at the preprocessing stage. For example, the terms product and industry under topic 1 and 2 respectively are not as informative as the other terms in the group and can be eliminated from the vector model.

Block 5: bottom-up ontology extension

Ontologies play a vital role in manufacturing knowledge management. One of the main challenges related to development of ontologies is timely extension and evolution of the ontology such that they reflect the changes in the domain knowledge. Manual extension of ontologies is not efficient in the manufacturing domain where the body of knowledge is constantly evolving and new terms and concepts are introduced as a result of technological advancements. Development of automated tools that can support bottom-up ontology extension is a necessity.

Block 5 in the proposed method provides a semi-automatic approach for collecting the key concepts and their inter-relations that can be incorporated in formal manufacturing ontologies. After documents are categorized based on the discovered topics in Block 4, the collected terms under each topic can be first added to a thesaurus before being imported to the ontology. Figure 19 shows how the terms listed under Topic 2 are added as Candidate Concepts to a thesaurus that uses SKOS (Simple Knowledge Organization System) syntax and semantics.

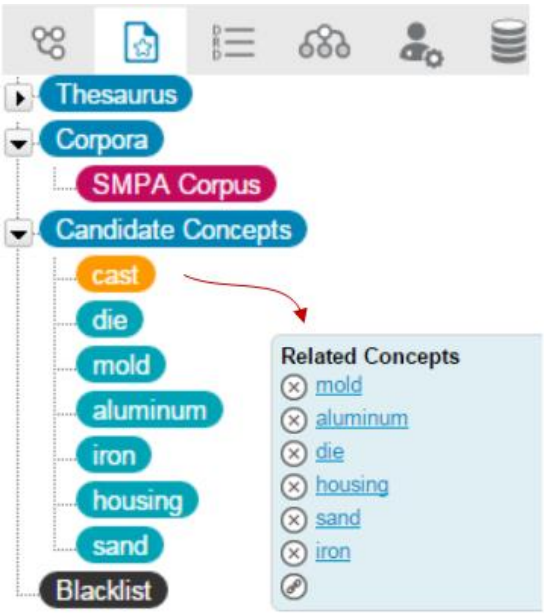


Figure 19. Extracted terms from topic modeling process can be imported to a thesaurus

Because the extracted concepts belong to the same topic, it can be inferred that they are semantically related. For this reason, they are connected to each other as Related Concepts in the thesaurus. Each topic can also be treated as a Concept Scheme, or group of similar concepts, in the thesaurus. The developed thesaurus serves as the steppingstone for a more formal and axiomatic ontology.

An Example for Pattern Discovery in Documents via Hybrid Method

As mentioned earlier, beside unsupervised classification of documents, one other advantage of proposed hybrid method is discovering possible patterns in documents. Patterns means a set of primary or secondary topics which are possibly discussed in documents. Patterns cannot be discovered by document classification since this technique needs predefined classes of training data to classify unknown documents under predefined classes. For instance, one may use document classification technique to make a distinction between Manufacturing-related document and environmental-related

document. However, by running the proposed hybrid method, the feasible patterns in each and every document become visible for the user. Example below can better clarify this idea. As it is mentioned in the previous section, cluster B was selected for topic modeling and user has determined two topics to be discovered within this cluster. As it is depicted in Tables 16 and Table 17, these two topics are related to machining and casting services. Now in this scenario, user desires to discover three topics or patterns (N=3) instead of two. Therefore, hybrid algorithm is again initiated to find three topics in cluster B. Table 18 shows these discovered three topics and their 10 most frequent terms.

Table 18. Top 10 stemmed terms in Topic 1, Topic 2 and Topic 3

	Topic 1	Topic 2	**Topic 3**
1	cast	turn	machin
2	aluminum	cnc	product
3	die	servic	custom
4	mold	tool	precis
5	part	part	process
6	rang	industri	manufactur
7	design	steel	cnc
8	steel	equip	mill
9	well	chuck	assembl
10	sand	materi	qualiti

This time, four first documents of cluster B are assigned to these three discovered topics. Table 19 lists the returning topic probabilities associated with each document.

Table 19. Documents and their topic probabilities

Document	**Topic 1**	**Topic 2**	**Topic 3**
1.xml	0.270332	0.387171	0.342497
2.xml	0.093007	0.697895	0.209097
3.xml	0.510078	0.124031	0.365891
4.xml	0.498575	0.195157	0.306268

The comparison between Tables 16 to 19 provides interesting information; Topic 2 (CNC machining-related terms), shown in Table 17 in the previous section, is

approximately a superclass of two other patterns. These two patterns are illustrated in Table 18 and are named Topic 2 and Topic 3. By interpreting Table 19, user can recognize that Topic 2 is related to the *Turning* process and Topic 3 is related to *Milling* process. The probabilities of each topics in a document determine the patterns which exist in that document. For instance, document 2.xml which refers to a manufacturer website at http://www.hammondmachine.com/cnc-turning-services.html , according to the Table 19 has patterns of *0.093 percent casting*, *0.697 percent turning* and *0.209 percent milling (0.093, 0.697, 0.209)*. However, it is worth to mention that the number of topics needs to be determined carefully by the user and through try and error processes. Up to this date, there is not any criteria to find the right number of topics since it mainly depends on the nature of dataset which is under examination.

Use Case Theorization for Manufacturing Self-Assessments

It is crucial for small to medium size manufacturers to keep up with new technologies and trends if they want to survive in a competitive manufacturing market. The advances in manufacturing domain are triggered by diverse technologies and processes such as 3D printing, CAD/CAM and etc. Therefore, product manufacturing companies, to not to be left behind other competitors, must adapt themselves to these rapid changes in technology and refine their industrial processes and offer new services. A good example for neglecting advances in industry is Nokia manufacturing company. They were producing high quality cellphones, however, they did not cop up with new trends and they lost the market. Manufacturing companies, irrespective to their size and capacity, need to plan new strategies to address these changes in order to capture their growth and subsequently their survival. Therefore, a successful manufacturing company

need to conduct periodic self-assessments to evaluate and adapt itself in conjunction with other competitors and market needs. The above-mentioned Hybrid method can help companies to search new processes, services, facilities and many more by modeling their competitors based on the extraction of the most important topics or concepts in which they are describing in their websites. As a result, they can take advantage of these topics to be inspired and adapt or innovate themselves in a disruptive environment.

Summary

As the collective manufacturing knowledge is further transferred into the web in the form of unstructured text, it becomes more difficult to find relevant information from web-based resources to inform various decision-making activities throughout product value chain. There is a need for new computational tools to help organize and search this vast data and extract and understand the knowledge patterns hidden in the data. In this chapter, a novel approach for facilitating search and organization of textual documents and also extraction of thematic patterns in manufacturing corpora using document clustering and topic modeling techniques is proposed. Topic modeling is a powerful technique for classifying and characterizing hidden themes in document corpora. The proposed method adopts LDA algorithm for application of topic modeling. Manufacturing supplier discovery is used as the area of application and the webpages of contract manufacturing service providers are used to build the corpus. Using topic modeling in conjunction with document clustering facilitates automated annotation and classification of manufacturing webpages, thus improving the intelligence of supplier search and information retrieval tools. Furthermore, the extracted terms from the topic

modeling process can be integrated with different reference models such as manufacturing thesauri and ontologies to enable bottom-up knowledge elicitation.

IV. CONCLUSION AND FUTURE WORK

The objective of this research was to use different text mining techniques in order to analyze manufacturing textual data for preliminary evaluation and classification of manufacturing suppliers. A set of research questions was identified in Chapter 1. To answer the identified questions, different text mining and machine learning techniques were developed, implemented, and tested throughout this study. In this chapter, the findings related to the identified research questions are summarized. In addition, the main contributions of this work, together with the proposed future works are outlined in this chapter.

Answers to Research Questions

1. How supervised text classification techniques can help rapid configuration of supply chain in manufacturing domain?
    - As it is discussed in chapter 1, one of the main challenges in rapid configuration of agile supply chains is finding the right suppliers who possess the required set of technological capabilities and competencies. The relationship between the participants of the agile supply chain, particularly at the early stages of supply chain formation, is often virtual and based on electronic and web search interactions. It is discussed in chapter 2 that websites of manufacturing suppliers are rich sources of data which should be explored thoroughly. These data are big in size and quantity and are needed to be classified by the customers before they choose their most appropriate suppliers which meet their requirements. The first action a customer may take for its supplier discovery is through a web search so they can get access to the

profiles of the suppliers. However, the basic result of a web search only pertains to the limited vocabularies used in the search and does not provide detailed classification of the suppliers. Therefore, the human expert needs to use their judgment to classify a supplier. This is a very tedious task for even human experts which can increase risk of neglecting a candid supplier. To avoid this issue, an automated and supervised text classification technique can be used for rapid categorization of suppliers based on their capabilities in services they offer. To this end, a text classification technique based on Naïve Bayes method for classification of the providers of manufacturing service based on their capability narratives. This automated classification technique enhances the intelligence and efficiency of service discovery, evaluation, integration, and composition process in any service-oriented manufacturing environment.

2. How the performance of the text mining techniques in manufacturing domain can be improved?

   In this work, four strategies are developed to improve the performance of text mining techniques in the manufacturing domain. Each of these strategies are described below:

   - The first common step in any supervised or unsupervised text mining task, whether it is classification, clustering or topic modeling, is to extract data from its resources and transforming it to a convenient format which is most suited for the applications that processing it. Due to the nature of manufacturing technology which is constantly changing and evolving, training

data are always changing and they need to be updated and archived for future text analytics tasks. Chapter 2 recommends to import data extracted from manufacturing websites in XML format. In this way, all training or test data can be archived and tagged with appropriate XML tags such that they can contain important metadata such as date of extraction, process name, corresponding manufacturers and many more.

- The "tm" package in R programming language takes advantage of an external dictionary of general terms known as stop words. Generally, these words are common in any literature. Since these words are less informative, they have to be removed during preprocessing steps. However, every literature has its own vocabularies which are considered as common terms within that domain. Chapter 2 suggests a list of terms which can be extendable and are considered as general terms used in the manufacturing. These terms should be removed during preprocessing techniques. This action reduces the dimension of DTM matrix introduced in chapter 2 and consequently, improves the performance of text mining tasks in manufacturing domain.
- As discussed in chapter 2, text mining techniques, such as classification and clustering, can only process single terms and are not capable of accepting compound phrases. On the other hand, breaking down compound phrases will result in loss of meaning. To resolve this flaw, the most important compound phrases related to the processes of interest and available in the corpus should be replaced by their abbreviations. Therefore, compound terms are treated and processed as single terms which is suitable for text mining tasks. For example,

"Electrical Discharge Machining" is substituted with EDM during pre-processing. This improves the performance of text classification process since not only the dimension of DTM matrix is reduced, but also the compound terms maintain their meaning as a single terms.

- The performance of supervised machine learning and text classification highly depends on the quality of the training data. In chapter 2, two different approaches are used for preparing training data or dictionary of the terms, namely, term-based method and concept-based method. The term-based method, is a standard method which uses a known set of textual capability narratives from corpus to form the training data. In the concept-based method, a formal thesaurus called ManuTerms is used to create the training data through semantic querying of the thesaurus. Both methods were evaluated experimentally. It was observed that the accuracy of the term-based classifier increases as the number of training documents increase. However, collecting clean and high-quality training documents in sufficient quantity presented to be a challenge. The difficultly related to collecting training documents is further exacerbated when dealing with complex classes that require highly specialized documents. In Chapter 2 a concept-based method was proposed that utilizes a formal thesaurus, called ManuTerms, to provide a solution to this problem. ManuTerms is developed by domain experts in a collaborative fashion and, therefore, it is rigorously tested and validated and can be used as a precise source of highly relevant manufacturing terms. Furthermore, since it is based on SKOS, it can be linked to various datasets on linked data, thus

enabling its continuous and dynamic evolution and extension using crowdsourcing services. The main advantage of concept-based classification method is that the terms in the dictionary maintain their semantic connection with the key features of the classes of interest. Through conducting an experiment, it was demonstrated that the concept-based classifier performs considerably better when classifying documents under relatively similar and dissimilar classes. Also, it is less robust to the noisy environment of large textual corpora. In summary, the concept-based method proved to be more accurate, more robust, and less expensive compared to the term-based method.

3. How clustering and topic modeling can add extra values to rapid supply chain configuration?

    - Hybrid clustering and topic modeling introduced in chapter 3, help build clusters of manufacturing websites and discover the hidden patterns in the identified clusters. Furthermore, it harvests the key manufacturing concepts that can be imported to manufacturing thesauri and ontologies. In addition, the proposed hybrid method due to its unsupervised nature does not need any training data. This significantly reduces the initial setup cost and time. The reason is that, text classification classifies each document under only a predefined class and training data for that class have to be prepared before text classification tasks starts. However, in the introduced hybrid method, after categorizing documents (i.e., supplier webpages) under different clusters, topic modeling method is used to further classify each documents of the desired clusters under more than one category with known properties.

Therefore, secondary topics discussed in documents are readily highlighted. Moreover, when highly informative terms are clustered together under a topic, the likelihood of discovering interesting patterns in data increases.

Thesis Contributions

The main contributions of this thesis work are summarized as below.

- Investigating the application of text mining and machine learning techniques for classification and characterization of manufacturing suppliers for the first time.
- Development of a reusable manufacturing text corpus for text mining purposes.
- Introducing needs for specifically designed stopwords for manufacturing domain.
- Discussing importance of concatenation of compound terms in order to preserve semantic of natural language.
- Demonstrating concept-based datasets as a reliable source of training data for supervised classification in manufacturing domain.
- Introducing a hybrid method based on clustering and topic modeling for unsupervised and multi-classification of documents.

Future Works

The future work in regard to the chapter 2 includes extending and enriching ManuTerms systematically through mining the classified documents using the hybrid clustering and topic modeling technique which is introduced in chapter 3.

There are multiple areas that can be further explored in chapter 3. One future task is to evaluate the performance of different topic modeling algorithms that can be used in the proposed framework. In the current implementation, the number of topics is determined upfront by the user, but there is a need for calculating the optimum number of topics in the corpus automatically. The pre-processing step results in decomposing compound phrases into single terms. This may cause information loss and semantic degradation. Auto discovering of the compound phrases and concatenating them in the pre-processing step to preserve the semantic relevance of the terms defines another future research direction. In addition, the results provided in chapter 3 are only based on a single run of the mining process. The performance of the proposed method can be further improved through multiple iterations and subsequent elimination of less informative terms and concepts. Moreover, the corpus used in this proof-of-concept implementation only contained 100 documents. To reap the true benefits of text mining in manufacturing, the size of the corpus has to be increased significantly.

APPENDIX SECTION

APPENDIX A: An example of unknown Manufacturing document, classified as a CNC machining document through Topic Modeling technique.

```
<?xml version="1.0" encoding="UTF-8"?>
<Info>
<Supplier> Astro Co</Supplier>
<ExtractionDate> 03/08/2016 </ExtractionDate>
<Type> Manufacturing Document</Type>
<text>
Custom Machine Building
CNC Precision Machining
Welding and Fabrication
Panel Wiring and Control Systems Reverse Engineering
Machine Repair, Rebuilding and Refurbishment
5-Axis Machining and Milling
Our products are positively impacting companies worldwide in a select
group of industries, including:
Medical
Pharmaceutical
Energy
Food Processing
Government
Aerospace
Electronics/Semiconductor
Telecommunications
Packaging
General Manufacturing
CNC Precision Machining
Precision Machining Services
Astro Machine Works, Inc. provides a wide range of precision machining
services to meet your unique
parts requirements.
We specialize in single-part prototyping, short to medium production
runs,
and special blanket order arrangements with periodic scheduled
releases.
We provide precision machining solutions for a wide range of clients,
including
those in the medical, government, energy, aerospace and manufacturing
industries.
CNC Precision Machining Astro Machine Works features a wide range of
precision machining capabilities,
including a full complement of
computerized and manual milling and turning centers, as well as wire-
EDM and water jet and laser cutting capabilities.
Our CNC lathe machining capabilities offer small to very large
capacities.
The CNC lathes we use are custom configured for the most intricate
applications
and integrate with all of our CAD/CAM software systems.
Our CNC machining department, equipped with CAD/CAM software, has the
ability
```

to download customer drawing files from anywhere in the world, greatly streamlining our machining processes.
This has given us the ability to provide our customers with shorter turnaround times
and a lower cost-per-piece than was possible with
previous technology.
In addition to the CNC lathe capabilities offered to our customers, we also feature a robust offering of simultaneous 5-Axis machining and five-sided milling capabilities.
These machines allow our customers to obtain parts with complex shapes and geometries at a competitive price point.
This type of precision machining is also incredibly accurate, and provides a smooth surface finish when
it is needed most.Our Continued Investment in Precision Machining
At Astro Machine Works, Inc., our passion is efficient and consistent machining. We constantly strive
to provide our customers
with excellent value and trusted service, and doing that means staying up-to-date on emerging technology in our industry.
To ensure that we keep pace with our customers' needs, we have a formal capital
investment policy that mandates continued investment in computerized machine tools and
other state of the art technologies.
One example of this is our continued investment
into Electrical Discharge Machining (EDM) technology. By utilizing the unique benefits of the EDM machining process,
our customers are able to request smaller,
highly intricate or delicate pieces and receive them free of burrs and defects.
By utilizing methods and processes that reduce the number of setups, we are able to pass the increased
efficiency along to the customer
in the form of faster turnaround and cost-savings.Contact Us Today to Learn More
Astro Machine Works is dedicated to smooth, highly efficient precision machining.
Whether it's machining complex components, jigs, fixtures, or gages, or repairing or modifying parts,
Astro has the expertise to deliver the results you expect.
</text>
</Info>

REFERENCES

Agarwal, V., Thakare, S., & Jaiswal, A. (2015). Survey on Classification Techniques for Data Mining. *International Journal of Computer Applications*, *132*(4), 13-16. doi:10.5120/ijca2015907374

Aggarwal, C. & Yu, P. (2009). On clustering massive text and categorical data streams. *Knowledge And Information Systems*, *24*(2), 171-196. doi:10.1007/s10115-009-0241-z

Alghamdi, R. & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal Of Advanced Computer Science And Applications*, *6*(1). doi:10.14569/ijacsa.2015.060121

AlSumait, L., Barbará, D., & Domeniconi, C. (2008). On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. *2008 Eighth IEEE International Conference On Data Mining*. doi:10.1109/icdm.2008.140

Ameri, F., Kulvatunyou, B., Ivezic, N., & Kaikhah, K. (2014). Ontological Conceptualization Based on the SKOS. *Journal Of Computing And Information Science In Engineering*, *14*(3), 031006. doi:10.1115/1.4027582

Ameri, F., Urbanovsky, C., & McArthur, C. (2016). A Systematic Approach to Developing Ontologies for Manufacturing Service Modeling. *Proc. 7Th International Conference On Formal Ontology In Information Systems (FOIS 2012), Graz, Austria*.

Bérenguer, C., Grall, A., & Soares, C. (2012). *Advances in safety, reliability and risk management*. Boca Raton: CRC Press.

Bijalwan, V., Kumar, V., Kumari, P., & Pascual, J. (2014). KNN based Machine Learning Approach for Text and Document Mining. *IJDTA*, *7*(1), 61-70. doi:10.14257/ijdta.2014.7.1.06

Blei, D. (2012). Probabilistic topic models. *Communications Of The ACM*, *55*(4), 77. doi:10.1145/2133806.2133826

Chakraborty, G., Pagolu, M., & Garla, S. *Text mining and analysis*.

Chen, H., Fuller, S., Friedman, C., & Hersh, W. Knowledge Management, Data Mining, and Text Mining in Medical Informatics. *Medical Informatics*, 3-33. doi:10.1007/0-387-25739-x_1

Choudhary, A., Harding, J., & Tiwari, M. (2008). Data mining in manufacturing: a review based on the kind of knowledge. *J Intell Manuf*, *20*(5), 501-521. doi:10.1007/s10845-008-0145-x

Dong, B. & Liu, H. (2006). Enterprise Website Topic Modeling and Web Resource Search. *Sixth International Conference On Intelligent Systems Design And Applications*. doi:10.1109/isda.2006.25

Frigui, H. & Nasraoui, O. (2004). Simultaneous Clustering and Dynamic Keyword Weighting for Text Documents. *Survey Of Text Mining*, 45-72. doi:10.1007/978-1-4757-4305-0_3

Gantz, J. & Reinsel, D. (2011). Extracting Values from Chaos. *IDC'S Digital Universe Study, Sponsored By EMC.*, (IDC 1142), 1-12.

Gerber, M. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, *61*, 115-125. doi:10.1016/j.dss.2014.02.003

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings Of The 22Nd Annual International ACM SIGIR Conference On Research And Development In Information Retrieval - SIGIR '99*. doi:10.1145/312624.312649

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2013). Interactive topic modeling. *Mach Learn*, *95*(3), 423-469. doi:10.1007/s10994-013-5413-0

Khajehzadeh, N., Postelnicu, C., & Lastra, J. (2012). Detection of abnormal energy patterns pointing to gradual conveyor misalignment in a factory automation testbed. *2012 IEEE International Conference On Systems, Man, And Cybernetics (SMC)*. doi:10.1109/icsmc.2012.6377899

Korde, V. (2012). Text Classification and Classifiers:A Survey. *International Journal Of Artificial Intelligence & Applications*, *3*(2), 85-99. doi:10.5121/ijaia.2012.3208

Ku, Y., Chiu, C., Zhang, Y., Chen, H., & Su, H. (2014). Text mining self-disclosing health information for public health service. *Journal Of The Association For Information Science And Technology*, *65*(5), 928-947. doi:10.1002/asi.23025

Kung, J., Lin, J., & Hsu, Y. (2015). Using Text Mining to Handle Unstructured Data in Semiconductor Manufacturing. *Joint E-Manufacturing And Design Collaboration Symposium (Emdc), International Symposium On Semiconductor Manufacturing (ISSM), (IEEE, Piscataway, NJ, USA)*, 1-3.

Lau, R., Liao, S., Kwok, R., Xu, K., Xia, Y., & Li, Y. (2011). Text mining and probabilistic language modeling for online review spam detection. *ACM Trans. Manage. Inf. Syst.*, *2*(4), 1-30. doi:10.1145/2070710.2070716

Liu, Y., Loh, H., Toumi, K., & Tor, S. (2008). A hierarchical text classification system for manufacturing knowledge management and retrieval. *International Journal Of Knowledge Management Studies*, *2*(4), 406. doi:10.1504/ijkms.2008.019749

Loging, W. *Bioinformatics and computational biology in drug discovery and development*.

Lu Murphey, Y. (2015). Vehicle Fault Diagnostics Using Text Mining, Vehicle Engineering Structure and Machine Learning. *International Journal Of Intelligent Information Systems*, *4*(3), 58. doi:10.11648/j.ijiis.20150403.12

M. Blei, D., Y. Ng, A., & I. Jordan, M. (2003). Latent Dirichlet Allocation. *Journal Of Machine Learning Research*, *3*, 993-1022.

MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and

Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967.

M. Kornfein, M. & Goldfarb, H. (2007). A Comparison of Classification Techniques for Technical Text Passages. *Proceedings Of The World Congress On Engineering*, *2*, 1072-1075.

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

Martens, D. & Provost, F. (2014). Explaining Data-Driven Document Classifications. *MIS Quarterly*, *38*, 73-99.

Masseroli, M., Chicco, D., & Pinoli, P. (2012). Probabilistic Latent Semantic Analysis for prediction of Gene Ontology annotations. *The 2012 International Joint Conference On Neural Networks (IJCNN)*. doi:10.1109/ijcnn.2012.6252767

Maynard, D., Peters, W., d'Aquin, M., & Sabou, M. (2007). Change management for metadata evolution. *International Workshop On Ontology Dynamics - IWOD 2007, 7 Jun 2007, Innsbruck, Austria*.

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *FNT In Information Retrieval*, *2*(1–2), 1-135. doi:10.1561/1500000011

Pant, G. & Srinivasan, P. (2005). Learning to crawl. *ACM Transactions On Information Systems*, *23*(4), 430-462. doi:10.1145/1095872.1095875

Qi, X. & Davison, B. (2009). Web page classification. *CSUR*, *41*(2), 1-31. doi:10.1145/1459352.1459357

Sanchez-Pi, N., Martí, L., & Garcia, A. (2014). Text Classification Techniques in Oil Industry Applications. *Advances In Intelligent Systems And Computing*, 211-220. doi:10.1007/978-3-319-01854-6_22

Sebastiani, F. (2002). Machine learning in automated text categorization. *CSUR*, *34*(1), 1-47. doi:10.1145/505282.505283

Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, & Jiajun Bu, et al. (2014). Interpreting the Public Sentiment Variations on Twitter. *IEEE Trans. Knowl. Data Eng.*, *26*(5), 1158-1170. doi:10.1109/tkde.2013.116

Singh, N. (2012). Financial Statement Fraud Detection using Text Mining. *International Journal Of Advanced Computer Science And Applications*, *3*(12). doi:10.14569/ijacsa.2012.031230

SKOS Simple Knowledge Organization System Reference. (2016). *W3.org*. Retrieved 25 May 2016, from http://www.w3.org/TR/skos-reference

Spohrer, J., Maglio, P., Bailey, J., & Gruhl, D. (2007). Steps toward a science of service systems. *Computer*, *40*(1), 71-77. doi:10.1109/mc.2007.33

Steinbach, M., Karypis, G., & Kumar, V. (2000). A Comparison of Document Clustering Techniques. *KDD Workshop On Text Mining*, *400*, 525-526.

Steyvers, M. & Griffiths, T. (2005). Probabilistic Topic Models. *In T. Landauer, D Mcnamara, S Dennis, And W. Kintsch (Ed),Latent Semantic Analysis: A Road To Meaning, Laurence Erlbaum*, 1-15.

Ting, S., Ip, W., & Tsang, A. (2011). Is Naïve Bayes a Good Classifier for Document Classification?. *International Journal Of Software Engineering And Its Applications*, 5(3), 37-46.

Ur-Rahman, N. & Harding, J. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems With Applications*, *39*(5), 4729-4739. doi:10.1016/j.eswa.2011.09.124

Yang, T., Torget, A., & Mihalcea, R. (2016). Topic modeling on historical newspapers. *In Proceedings Of The 5Th ACL-HLT Workshop On Language Technology For Cultural Heritage, Social Sciences, And Humanities*, 96–104.

Zhai, Z., Liu, B., Xu, H., & Jia, P. (2011). Constrained LDA for Grouping Product Features in Opinion Mining. *Advances In Knowledge Discovery And Data Mining*, 448-459. doi:10.1007/978-3-642-20841-6_37