THE BUSINESS VALUE OF TEXT MINING.

Bachelor Degree Project in Informatics
Level 30 ECTS
Spring term 2017

Richard Stolt

Supervisor: Jeremy Rose
Examiner: Eva Söderström
Acknowledgements

I would like to give my thanks to the people I have had the pleasure of working with at iP.1 and for letting me conduct my research at the company. Leif Alvarsson, Mikael Linder and Erik Wikström surprised me by their availability and level of support, putting me in contact with the right people, and lending their time of the day to help me with the technical aspects.

I would also like to thank my examiner, Eva Söderström for her valuable comments and the extra set of eyes on the thesis. Though it is at first hard to admit to yourself the feedback as being necessary, the end results do in all aspects end in something that is better.

I would consider myself especially indebted to Jeremy Rose, my supervisor. Jeremy is not only a great teacher, but helped change my attitude toward scientific research, as it was something I initially wanted finished as soon as possible, to something that was motivating and enjoyable.
Abstract

Text mining is an enabling technology that will come to change the process for how businesses derive insights & knowledge from the textual data available to them. The current literature has its focus set on the text mining algorithms and techniques, whereas the practical aspects of text mining are lacking. The efforts of this study aims at helping companies understand what the business value of text mining is with the help of a case study. Subsequently, an SMS-survey method was used to identify additional business areas where text mining could be used to derive business value from. A literature review was conducted to conceptualize the business value of text mining, thus a concept matrix was established. Here a business category and its relative: derived insights & knowledge, domain, and data source are specified. The concept matrix was from then on used to decide when information was of business value, to prove that text mining could be used to derive information of business value.

Text mining analyses was conducted on traffic school data of survey feedback. The results were several patterns, where the business value was derived mainly for the categories of Quality Control & Quality Assurance. After comparing the results of the SMS-survey with the case study empiricism, some difficulties emerged in the categorization of derived information, implying the categories are required to become more specific and distinct. Furthermore, the concept matrix does not comprise all of the business categories that are sure to exist.

Keywords: Text Mining, business value, business value of text mining, survey data analysis
# Table of Contents

1. Introduction .................................................................................................................. 5  
   1.1. Aim and Objectives ................................................................................................. 5  

2. Background .................................................................................................................... 8  
   2.1. Text Mining ............................................................................................................ 8  
       2.1.1. Information Retrieval & Information Extraction .............................................. 8  
       2.1.2. Topic Tracking ............................................................................................... 8  
       2.1.3. Text summarization ....................................................................................... 9  
       2.1.4. Categorization ............................................................................................. 9  
       2.1.5. Clustering .................................................................................................... 9  
       2.1.6. Association Rule Mining .............................................................................. 10  
       2.1.7. Opinion Mining & Sentiment Analysis ....................................................... 10  
   2.2. The Business Value of Text Mining ....................................................................... 11  
       2.2.1. Defining the business value of Text Mining .................................................. 11  
       2.2.2. Quality Control and Quality Assurance ....................................................... 13  
       2.2.3. Customer Relationship .............................................................................. 14  
       2.2.4. Comprehensive Summaries of Text ............................................................. 15  
       2.2.5. Examples of information derived that are not of business value ............... 16  
       2.2.6. A Concept Matrix explaining the Business Value of Text Mining .............. 16  

3. Problem definition ......................................................................................................... 19  
   3.1. Problem statement ................................................................................................. 19  
   3.2. Limitations of the study. ....................................................................................... 20  
   3.3. Expected outcome ............................................................................................... 21  

4. Method .......................................................................................................................... 22  
   4.1. Text mining case study on dataset of survey feedback ....................................... 22  
   4.2. Participants ........................................................................................................... 23  
   4.3. Mini-literature review ......................................................................................... 24  
   4.4. SMS-survey method ......................................................................................... 24  
   4.5. Analysis ............................................................................................................... 25  
   4.6. Research Ethics .................................................................................................... 26  
       4.6.1. Anonymity and Confidentiality ..................................................................... 26  
       4.6.2. Disclosure .................................................................................................... 27  

5. Research execution ....................................................................................................... 28  
   5.1. Mini-literature review ......................................................................................... 28
5.2. TM Data source selection ............................................................................. 28
  5.2.1. Choice of Data ....................................................................................... 29
5.3. Text Mining the Dataset ............................................................................. 30
  5.3.1. Analysis of TM results .......................................................................... 31
5.4. SMS-survey ................................................................................................. 33
  5.4.1. SMS-survey analysis ............................................................................ 34
6. Analysis ........................................................................................................... 35
  6.1. Text Mining ................................................................................................. 35
    6.1.1. Finding patterns by N-grams ............................................................... 35
    6.1.2. Correlation analysis of found patterns ................................................. 37
    6.1.3. Validate findings using the raw data .................................................. 38
    6.1.4. Sentiment Mining ............................................................................... 42
  6.2. SMS-survey ................................................................................................. 46
    6.2.1. Nominal responses .............................................................................. 46
    6.2.2. Qualitative responses ......................................................................... 47
7. Results ............................................................................................................. 48
  7.1. Information of Business Value .................................................................... 48
    7.1.1. Information of Business Value derived ................................................ 48
    7.1.2. Learning from Dataset ....................................................................... 49
  7.2. SMS-survey ................................................................................................. 50
8. Discussion ......................................................................................................... 51
  8.1. Reflections on the research approach .......................................................... 51
  8.2. Discussion of results and recommendations in relation to the traffic school 52
  8.3. Meta-Analysis of the TM-methods to derive information of business value 53
    8.3.1. Toward a hypothesis for the business value of TM ............................. 56
  8.4. Contributions to Text Mining research ....................................................... 57
  8.5. Scientific aspects ......................................................................................... 57
  8.6. Socio-ethical aspects .................................................................................. 58
9. Conclusion and future research ..................................................................... 60
10. References ...................................................................................................... 62
1. Introduction

Studies indicate that 80% of a company’s information is contained in text documents (He, Zha & Li., 2013; Tan, 1999). In regards to big data, recent sources give indications of 5% of the data being structured (Cukier 2010), whereas, 95% is unstructured (Gandomi & Haider 2015). Additionally, unstructured data are not only text documents, they are also of formats such as video, image, and audio; therefore, often lacking the traditional structure and organization required by machines for analysis (Gandomi & Haider 2015). Employing the means of extracting insights and knowledge from such a source could prove to be of significant value to a business.

Text mining (TM) attempts finding meaningful patterns in unstructured data. The data are usually originating from unstructured text (Fuller, Biros & Delen., 2011). Other works define it as being focused on finding and extracting meaningful information, knowledge, non-trivial patterns, models, directions, trends or rules from unstructured text documents. (Abdous & He, 2011; Feldman & Dagan, 1995; He, Zha & Li., 2013; Hung & Zhang, 2011; Tan, 1999).

The business value of using text mining (TM) for making sense of data grow apparent when appearing in larger sizes. Extracting information is harder for humans as the quantity of text grows. Reading only a few sentences or messages out of many for decision-making, may lead to a biased view (Hu & Liu, 2004). The study, therefore, focus on the business value of TM that is derived when the human is no part of the earlier stage, manually making sense of numerous texts. Evidently, the literature present TM as a technology of business value (e.g. He, Zha & Li, 2013), however, there is little research on the topic conducted with a purpose of demonstrating the business value of TM. In a literature review by Melville, Kraemer & Gurbaxani (2004), they learned that Information Technology (IT) “is valuable, offering an extensive menu of potential benefits ranging from flexibility and quality improvement to cost reduction and productivity enhancement.” On the subject of “business value of text mining” not much is offered in terms of research; however, the research conducted do reveal implications toward it being valuable in a business setting. To conduct an investigation of the business value of Text Mining is therefore an important direction of research. As there are still many areas to be explored, in regards to the implications of its business value, the study finds it necessary to gain understanding of text mining in a business setting, to learn more about its synergistic nature.

1.1. Aim and Objectives

Considering the abovementioned, the overall aim of the study is therefore:

“To conduct a purposeful investigation of the business value of TM”

The word purposeful investigation is used as a way to make it clear that the evidence of this study is set to be explicitly oriented toward the business value of TM, thus
differentiating from the rest of the research conducted prior to this study (cf. implicit evidence in the research mentioned in section 3; cf. discussion of the differences in section 8.1). To this end, the study entail and requires the initial objective of:

1. “To conduct TM analyses, ensuring the business value of TM on a general type of Dataset”

The type of datasets are not guaranteed to be originating from the same data source (e.g. Reviews, Surveys or Social Media etc.) or be the case that businesses are not active in the same business domains (e.g. companies active in different markets such as Hospitality or Manufacturing). In such circumstances it is therefore conclusive that datasets belongs to different contexts, such as different products or different services (this does mean that the empiricism of objective 1 is weak in this regard). The empirical evidence of objective 1 will investigate the synergistic nature of TM to the dataset. It follows that it is out of necessity to prove the business value of TM more generally, leading to the second objective:

2. “To conduct an empirical investigation of companies, identifying general problematic business areas where business value from TM could be derived”

The second objective investigates companies with data that are similar (further explanation of what similar in this contexts implies in section 4.5) to that which was used for the empirical evidence gathered from objective 1. The second objective will therefore investigate if TM can be synergistic to the data described by the companies that partake in the investigation. However, it is highly dependent on the response frequency of the participants in the investigation, in order to derive strong empirical evidence.

The final objective is:

3. “To compare the results from the business value as derived from TM the Dataset, with the identified business areas where such business value could be derived”

By a final comparison made on the empirical evidence gathered of the first objective 1, and the second objective 2, a stronger claim to the business value of Text Mining would be derived. However, given that the three objective are to some degree capable of being accomplished, giving an answer to their responding research questions (presented in section 3.1) will not be an issue.

The structure of the thesis is as follows. Section 2, present general theory for TM, and business value of TM. Section 3, introduce the problem of the established domain, and the inherent research questions, followed by, expectations and limitations of the study. Section 4, present the research method in detail. Section 5, present research execution. Section 6, analyses of text mining the dataset, and manual analysis of SMS-survey. Section 7, presenting the results of the prior analyses. Section 8,
discussion of the study. Section 9, conclusion and future research. Section 10, references.
2. Background

2.1. Text Mining

The subsection of TM, presents some of the more general methods for Text Mining in the literature. When it comes to the subject of Text Mining, much of the literature is too varied to give a clear description of how it is conducted and looks in its actuality. The latter meaning, the field of text mining is vast as its techniques varies, methods, differentiate, its dependability on who has conducted the research, and the data on which the research was conducted on. The field of Text Mining as will be presented in this section illustrates what TM is capable of by the different methods to attempt clarify how TM could be used for the extraction of information. Knowing how Text Mining can be used in different contexts, will help the reader understand different ways the technology can be of use.

2.1.1. Information Retrieval & Information Extraction

Information Retrieval (IR) could shortly be described as the gathering of, and search for, useful documents in a collection, and the indexing of text. It is an automated process, where all relevant documents are retrieved, simultaneously, mitigating the retrieval of non-relevant documents (Kosala & Blockeel 2000).

Information Extraction (IE) is a separate method, usually following the use of an IR system (Kosala & Blockeel, 2000). The goal is to transform data from being unstructured to structured, which is more easily digested and analyzed. It either processes unstructured or semi-structured data. The former type relies on linguistic pre-processing e.g. syntactic-, semantic-, and discourse analysis. The latter type in IE uses metadata, for this document that would intend author, date, and word count (Kosala & Blockeel, 2000). IE has two sub-tasks, Entity Recognition (ER) and Relation Extraction (RE). ER algorithms classifies text into predefined categories such as: person, date, and organization. RE algorithms identifies and extract semantic relationships for said entities. Extracting relations in a sentence such as “Adam Weishaupt (1748–1830), founder of the Bavarian Illuminati” would provide FounderOf[Adam Weishaupt, Bavarian Illuminati](Gandomi & Haider, 2015).

2.1.2. Topic Tracking

Topic tracking systems enables the tracking of documents (or categories) of interest, based on pre-specified or automatically predicted preferences (Gupta & Lehal, 2011). Topic tracking is applicable in circumstances where companies’ wants to be alerted of competitors’ or their own activities in news, keep up with competitive products or changes in the market. It can be used as a refinement step, with categorization or text summarization, on a volume of documents, as it could pre-specify the relevance of documents, based on keywords in their content, when searching for information on a topic (Gupta & Lehal, 2009).
2.1.3. Text summarization

Text summarization is used to convey key information from original text(s) in applications, such as, scientific and news articles, advertisements, emails and blogs (Gandomi & Haider, 2015). At its core, a summarization has the objectives of determining what the important parts of a text are, followed by, deciding how much of the content is to be reduced (Hahn & Mani, 2000). Reducing length and detail, while keeping a documents main points, is helpful when the end-user has to quickly judge the document relevancy and worth (Gupta & Lehal, 2009). A summary can indicate what sources are of relevance, give concise factual information, and give a critical opinion statement on content (Hahn & Mani, 2000). There are two different types of approaches to text summarization. In extractive approaches, a summary is a subset created from the original document. Representatives of sentences, the salient units of a text, are extracted and strung together. Text units are evaluated by analyzing frequency and location in text. It does not require understanding of the text. Abstractive approaches extract semantic information from text. Summaries are of text units not necessarily present in the original text (Gandomi & Haider, 2015). It is more a complex approach as it incorporates NLP techniques, lexical resources e.g. WordNet, and ontologies, resulting in more coherent summaries (Gandomi & Haider, 2015; Hahn & Mani, 2000). Extraction approaches are said to be more adaptable to large sources (Hahn & Mani, 2000), such as big data (Gandomi & Haider, 2015), as they identify certain segments of text, such as sentences, phrases or passages, which are mostly representative of the document’s content (Hu & Liu, 2004). Abstractive approaches are more sophisticated for the reason earlier stated, more coherent summaries, meaning the summary enriches the source content (Hahn & Mani, 2000).

2.1.4. Categorization

Categorization seeks to identify and classify the main theme of a document by placing said document into a pre-defined set of topics (Gupta & Lehal, 2009; Pang & Lee, 2008). The number of classes are dependent upon the complexity of a taxonomy, for example, in dealing with two classes (binary classification), to a thousand possible classes (Pang & Lee, 2008). It is different from IE, which has the aim of finding relations of entities or terms. Categorization use term frequency to count word appearance, and by enumerating the frequency of their appearance, judge or identify the main topics of said document (Gupta & Lehal, 2009). Considering the latter, it is worth noting that there is no actual information being extracted in comparison to IE.

2.1.5. Clustering

Clustering is a technique, which group related documents on the basis of some similarity measure e.g. distance metrics such as k-nearest neighbor (supervised) or k-means (unsupervised). The grouping is done automatically, without any pre-specified categorization; thereby differentiating from Categorization (Gupta & Lehal, 2009).
The technique is usually referred to as being unsupervised (Gupta & Lehal, 2009), but it depends on which specific technique is being adopted. It creates a vector of topics for each document and measures their weights on how well a given document fits into each cluster (Gupta & Lehal, 2009). Hung & Zhang (2011) used hierarchical clustering to create their root node of all documents on the topic “Mobile Learning”, followed by four sub clusters in which each corresponding document, leaf node, was divided into. Depending on the applied method, categorization and topic placing (naming them) can be done before or after such an identification, as done after placing them in clusters by Hung & Zhang (2011).

2.1.6. Association Rule Mining

Association rule mining is a technique which attempt finding relationships of variables in a given dataset (Gupta & Lehal, 2009). Netzer, et al., (2012) use an adapted technique of association rule mining, where they measured lifts to find the co-occurrence between different entities, and terms, from forum discussions. It discovers relationships by frequency of two or more recurring entities, in the same sentence or message (Gupta & Lehal, 2009). It can be used in business to see what items are typically purchased together, and derive a strategy for increased sales, from such information (Gupta & Lehal, 2009).

2.1.7. Opinion Mining & Sentiment Analysis

Opinion Mining and Sentiment Analysis are techniques which enables the analysis of opinionated text, toward entities such as products, organizations, individuals, and events (Gandomi & Haider, 2015). Pang & Lee (2008) refer to Dave, et al., (2003), saying, the ideal opinion mining tool would: “process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)”.

Accordingly, subsequent research fits this description, often emphasizing the analysis and extraction of various aspects on given items (Pang & Lee, 2008). In this report the terms are viewed as being synonymous, referring to both Opinion Mining and Sentiment Analysis, when either is mentioned.

There are several domains where the technology could be applied. Websites which solicit reviews e.g. Epinions, Amazon, or IMDB etc, are viable for its use, as a source to understand how products or services are perceived (Pang & Lee, 2008). Social media monitoring and analysis (He, Zha & Li, 2013) could be applied to monitor public relations, gain competitive intelligence, and track company image or product image (Pang & Lee, 2008). Considering the latter, Business Intelligence and Government Intelligence would seem to gain much value from SA. Pang & Lee (2008) give an example for how a computer manufacturer, now enabled by deductive technologies, is capable of answering why a certain laptop has unexpectedly low sales, and take action according to gathered information.
Pang & Lee (2008) describe the characteristics of document-level sentiment analysis as “a document can consist of sub-document units (paragraphs or sentences) with different, sometimes opposing labels, where the overall sentiment label for the document is a function of the set or sequence of labels at the sub-document level”. By analysis of sub-document units, and to explicitly utilize the relationships, one might achieve a more accurate global labeling (Pang & Lee, 2008). Gandomi & Haider (2015) divides the techniques of SA into three sub-groups: document-level, sentence-level, and aspect-based.

Document-level techniques determine if a given document express negative or positive sentiment. It assumes a document containing sentiment of a single entity (Gandomi & Haider, 2015). Sentence-level techniques attempt determining the polarity of a single sentiment about a known entity in a single sentence (Ganomni & Haider, 2015). Challenges of sentence-level techniques is to identify features and distinguishing the subjective (opinion) from objective (fact), in each sentence (Gandomi & Haider, 2015; Pang & Lee, 2008). Aspect-based techniques recognize all sentiments and aspects (feature or attribute) of an entity (e.g. product) in a document, and identify which sentiment is referring to what feature (Gandomi & Haider, 2015). In a product review of books, the entity would be book name, feature could be story, and sentiment is negative/positive. An example is, therefore, “The new Harry Potter (Title) has a great (Sentiment) story (feature) - Harry Potter: great story. Depending on the context of its use, valuable information about the different aspects of a product can be identified, and would be missed if only identifying sentiments (Gandomi & Haider, 2015).

### 2.2. The Business Value of Text Mining

This section presents the business value of TM as it appears in the literature; revealing different how TM methods and applications, in their relative contexts, lead to a certain business value being derived. First, the notion of the business value of text mining (as referred to in this study) is defined. Then the business value derived from Text Mining in the literature is presented and explained why it is relative to the context of business value of text mining. Subsequently, the concept matrix for the business value of text mining is created from the mini-literature review, and an explanation to how the concept matrix was established is made. Section 2.2.5 explains what is not considered to be business value of text mining.

#### 2.2.1. Defining the business value of Text Mining

In the context of management, business value is defined on Wikipedia (2017) as: “an informal term that includes all forms of value that determine the health and well-being of the firm in the long run.” The use of the term is too ambiguous and vague for a precise use in the context of this study. Oxford Dictionaries (2017), give a definition to the term “value”, as the following: “The regard that something is held to deserve; the importance, worth, or usefulness of something.” The author attribute information as being of business value, when it is considered to be “meaningful” (cf.
The aforementioned definition of text mining) or “useful”. Therefore, value is in this context not characterized as being purely monetary-based or to be mistaken for the concept of values (e.g. moral values).

The definition given by Someh & Shanks (2016) of business value, in the context of IT, state the business value to be generated from the use of emergent IT-enabled business systems. Emergent IT-enabled business systems thus meaning “informational IT and transactional IT systems and their complementary interactions” (Someh & Shanks, 2016). The emergent IT-enabled business system referring to the degree it is: “able to leverage analytical insights provided by the informational system and embedded in the transactional system” (Someh & Shanks, 2016). IT-enabled business systems generate transactional, informational, and strategic benefits (Someh & Shanks, 2016).

Transactional benefits include:

- Process efficiencies
- Effectiveness
- Cost reduction.

Informational benefits include:

- Fact-based decision making
- Real-time decisions
- Single version of the truth
- Actions based on facts

Strategic benefits include:

- Time to market
- Increased revenue
- Superior customer experience

Prior research in the field of business value of IT often seek to validate the causal relationships between IT and profit such as Lee (2001), and look at IT investments with what could be described as a top-down approach. The study at hand does not aim at investigating how profitable an investment in Text Mining technologies is in terms of costs. Instead, this study would be described as perceiving business value of text mining with a bottom-up view, where the emphasis is put at a lower level of abstraction, in this case, the informational benefits of an IT. To clarify on the latter, Someh & Shanks (2016) refers to Informational IT System Quality as being composed of assets and capabilities. Informational assets, the information-based assets which
enable analytical capability (e.g. data mining tools, OLAP). Informational capabilities, information include analytical human skills to analyze and generate insights from data, “management quality in planning, implementing and measuring initiatives, analytical processes and routines and analytical culture in the organization” (Someh & Shanks, 2016). This study argues that limiting the business value of IT to only be evaluated, when the data (return of investment data) is available to you after the investment is made does not help companies in the earlier stages of such an investment. Looking at the business value of an IT before the investment, is more important to know if it has synergy with the end-target in mind (the company who will derive value from the technology), it is also important to know if the end-target is actually a suitable prospect for a given IT investment. The latter is considered appropriate considering the fundamental use of text mining technologies is to find meaningful patterns in unstructured data. Therefore, the study view these patterns, if meaningful and thus informative, to be the core value (business value) that is supplied with the technology. To investigate its business value on a dimension (dimensions which implies costs) other than the informational should be considered an inappropriate approach.

Taking the abovementioned into consideration, the study use the following in-context definition and explanation of “The Business Value of Text Mining” to help apprehend the concept as the study will refer to it:

*The Business Value of Text Mining are the possible benefits, insights & knowledge, derived from applying the technologies of Text Mining in a business setting. To clarify, “benefits, insights & knowledge” refer to the valuable possible-information that is extracted when using a TM technology.*

Since the aims of this study encompass value of TM in a business setting, the information can and should be divided into its related category of business (categorization is explained in section 2.2.6). The following sections (section 2.2.2, section 2.2.3, section 2.2.4, and section 2.2.5) presents business value of TM in the literature, and the according information of business value that was derived.

### 2.2.2. Quality Control and Quality Assurance

The characteristics of the category “Quality Control and Quality Assurance” is the capability of verifying the quality of a product or service, and how well they currently match a set criteria, or requirements, of the customers, users, subscribers or developers. The benefits are discovering new means of improvements to a product or service, during or after its development in a business setting. A question in accord could be: “How do we improve a product/service?”

In a study by Abdous & He (2011) TM was applied to data generated from the input of student interactions, participating in live video streaming courses, with the purpose of improving the learning experience, in said courses. Mainly technical issues were identified in the messages sent, during a lecture, which were used to identify ways of
improving the learning experience. An interesting event in the study was how they, by review of text messages, identified a high frequency of requests for a full-screen option, which is reported to have been added into a recent update to the interface. The study demonstrates that a platform such as live video streaming courses, can derive value in applying TM.

Abrahams, et al., (2015) created a TM framework, to detect and discover defects in products. The TM method was applied to user generated content, in the automotive and consumer electronics domain, discovering defects through words (distinctive terms, product features, semantic factors gave the best result), in discussions and posts, on social media. The business value is derived as an early discovery of defects can help manufacturers in quality improvement, thus minimize selling and production of defective product units. Meaning, customer dissatisfaction is decreased as well as the warranties costs and defect-associated costs. A query for “exploding Samsung Note 7” gave 351,000 results on Youtube (February, 2017), referring to the infamous defective Samsung Note 7 event. Indicated as a number which probably could have been, or was, reduced by such a method.

Netzer, et al., (2012) gain insights and knowledge by applying TM on data gathered from online forums on discussions, related to Drugs and Cars. Revealing how different entities (brands) and related terms, by their frequency in the discussions, co-occur in the data. The analysis showcase the ability to quantify what consumers wrote about each car, being externally validated with survey data. The study show TM value in capable of specifying a sentiment e.g. “problem”, and its co-occurrence with terms like fuel, sludge or battery for the brand Toyota or any other brand. Applying a similar approach could be used to identify ways of quality improvement, by using the technology for quality control, identifying what aspects of a car (or product) that lead to customers’ disapproval.

Jurado & Rodriguez (2015) applied Sentiment Analysis techniques to identify and monitor underlying sentiments in text, written by developers in issues and tickets on well-known open Github-projects. The results is the capability of monitoring emotion such as surprise, anger, fear or sadness, on events or topics in a project. The findings give implications of benefits for use in large scale projects, to highlight issues needing to be addressed, for quality assurance.

2.2.3. Customer Relationship

The characteristics of what falls in the category of “Customer Relationship”, are the additions to customer service, communications, and customer insights or knowledge. Added benefits is the potential of discovering new means, with an impact on customer satisfaction in a business setting. A question could be: “How do we increase our customer satisfaction?”

He, Zha & Li (2013) give a demonstration of the business value in applying TM. Here, as an additional customer service and communication tool, to gain insight into
customers’ needs, wants, concerns and behaviors, in order to serve them better. As an investigative tool, TM is shown capable of supplying insights and knowledge from social media data, of a company. It is able to reveal the competitor strategy, by mining their Facebook or Twitter, for information on how customer engagement, promotion of services and customer bonding should be conducted.

Xiang, et al., (2015) applied TM to big data, from reviews in the hospitality domain. Insights and knowledge is gained, as TM reveal how the words, as they appear in a review and in what context, lead to certain ratings. The analysis therefore suggested, what the factors leading to better satisfaction, for a hotel experience, were. Here, for example, hygiene was an important factor to high satisfaction rating, and words related to free services e.g. breakfast airport shuttle. Knowing what factors tacitly lead to certain ratings is of high business value, as the knowledge lets a hotel owner know exactly what they could improve for better customer relationships (also service improvement).

Ikeda, et al., (2013) applied TM for user profiling based on the tweets written by a user. The result show TM as capable of making demographic estimations, by effectively estimating user occupation, age, area and hobby. Knowing whom one is selling to, and what their preferences are, is valuable information for various businesses such as, manufacturing businesses, sales businesses or service providers, etcetera.

2.2.4. Comprehensive Summaries of Text

The characteristics of the category “Comprehensive Summaries of Text”, is the outsourcing, automation, of activities such as manual analysis of large volumes of texts. The benefits are comprehensive summarized texts, presentation of information, reduced human biases, and reduction of redundant tasks. A question could be: “What is the general opinion of X-subject?”

The work by Hu & Liu (2004) used a variation of text summarization, named feature-based summary. Using WordNet, a semantic lexical database, sentiment analysis was conducted to identify if specific product features had opinions of negative or positive orientation. The study illustrates TMs effectiveness, being able in extracting sentiments from each review of a product, by using sentiment analysis on distinct sentences. The result was a summary of the products positive and negative sentiments, also displaying what feature of the product the sentiment was referring to.

Hu, Ko & Reddy (2014) used TM to gain insight into how purchase decision are made by customers on the web. The findings reveal sentiments having bigger impact on sales rank of products than ratings. It is implicated that the ratings of products are used as a screening device, followed by the decision being made after having viewed the sentiments. Having a summary of product sentiments, could prove to be highly valuable for the customers making purchase decisions on a larger amounts of
reviews. The same conclusion could be drawn for the perspective of the manufacturer, quickly able in gaining understanding of positive and negative aspects of a product.

Hung & Zhang (2011) applied a combination of TM and bibliometrics, on research abstracts, categorizing text on the topic of Mobile Learning, to find similarities of meaningful and content related words. The result of the study provide a quick, comprehensive and summative overview of a pre-specified scientific field, in this case Mobile Learning. The method found patterns, themes and trends, such as journals publishing preferences, frequency of topics over time, and topic predominance by country.

2.2.5. Examples of information derived that are not of business value

This subsection is text mining-derived information, which the author views as examples that does not fit the general notion of a business environment. It serves the purpose of showing contrast, to distinguish, help reduce ambiguity, by demarcating the less relative, when this study refers to “The Business Value of Text Mining”.

Fuller, Biros & Delen (2011) demonstrates a wide usage for TM. The authors attempted the use of TM techniques and tools for detecting deception and lies. Their sample was gathered from military bases; therefore, tested on real-world data, and actual crimes with severe consequences. The result show its potential in aiding the detection of lies in text, also the combination of text and data mining techniques showing a successful application on real-world data.

Choi, et al., (2013) accessed public data, The New York Times, to TM unknown articles and identify if its analyzed contents show relation to terrorism. Automated content analysis for supply of specific subject and topic articles, can aid a researcher in quickly finding the most relative articles when conducting a research.

Schumaker, Jarmoszko & Labedz., (2016) used social media data, i.e. twitter, to predict wins in the Premier League. Using sentiment analysis on crowdsourced sentiments, the proposed system prove it can be a better predictor of match outcomes than crowdsourced odds.

2.2.6. A Concept Matrix explaining the Business Value of Text Mining

A systematic approach to reviewing the literature has by identification, selection and extraction of the business value from each article, enabled the formation of three categories: Customer Relationship (CR), Quality Control and Quality Assurance (QC & QA), and Comprehensive Summaries of Text (CST). Furthermore, the author argues the necessity of creating categories, by adopting the definition for Categorization given by Jacob (2004) as the: “process of dividing the world into groups of entities whose members are in some way similar to each other. Recognition of resemblance across entities and the subsequent aggregation of like entities into categories lead the individual to discover order in a complex environment.” The result
is presented in an organizing framework in the form of a concept matrix (Bhattacherje 2012, p. 21), to illustrate how each article relate to a specific category (Observe Figure 1.). The framework most notably illustrate the value derived from TM and its befitting category, viable domains for TM, and where suitable data could be extracted from, as it is emerged in the body of research.

Considerations are made to some of the articles arguably being related to other categories, however, the purpose of the categorization is to easier grasp the concept of business value, in the current context. It is also worth noting some of the different derived categorical value possibly being interconnected, thereby, having an indirect impact on each other.

<table>
<thead>
<tr>
<th>Customer Relationship</th>
<th>Quality Control and Quality Assurance</th>
<th>Comprehensive Summaries of Text</th>
<th>Article</th>
<th>Benefits, Insights &amp; Knowledge</th>
<th>Domain/Setting</th>
<th>Data source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Abrous &amp; He (2011)</td>
<td>Technical issues, Defect Detection, Product Disapproval, Highlight in-project issues</td>
<td>Online Learning, Automotive &amp; Consumer Electronics, Cars, Development</td>
<td>Platform input, UGC: Forums, Social Media, Development-projects</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Jurado &amp; Rodriguez (2015)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hung &amp; Zhang (2011)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>He, Ko &amp; Reddy (2014)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Concept Matrix for Business Value as identified in the literature

The mini-literature review has revealed Text Mining being capable and valuable when applied in a business setting (Abdous & He, 2011; Abrahams, et al., 2015; He, Zha & Li, 2013; Hu & Liu, 2004; Hu, Ko & Reddy, 2014; Hung & Zhang 2011; Ikeda, et al., 2013; Jurado & Rodriguez, 2015; Netzer, et al., 2012; Xiang, et al., 2015). However, to the knowledge of the author, there is no research showing formal or general definition of the term, most notably is He, Zha & Li (2013), the only work in a text mining relative context, where the term is referred to informally (to the knowledge of the author); or Someh & Shanks (2016) though using the business value concept referring to IT generally.

The concept consists of three categories (also business areas) which are the following: Customer Relationship, Quality Control and Quality Assurance, and Comprehensive Summaries of Text. The information derived is divided and sorted into a suitable category, depending on the context and purpose of its end-use. (Observe Figure 1.). If information is of business value, depends upon the setting (domain) of its use, as some of the derived valuable information (i.e. benefits, insights & knowledge) could be as relative and of value in other non-business settings, e.g. identifying demographics could also be of interest in a research setting. Additionally, the use of the term business value is, in this study, distinguished from non-business settings, such as the military (Fuller, Biros & Delen, 2011) or when TM could be used for personal agendas, e.g. in the context of placing bets by predicting wins in the Premier
League using social media data (Schumaker, Jarmoszko & Labedz, 2016). The latter, non-business settings, could be argued for being the opposite, and surely there are more than those presented, however, for now they serve the purpose of grasping the concept, by demarcating that which is not considered business value, when using TM (i.e. the information is not relative to the current context and notion of what is considered a business setting).
3. Problem definition

The following section defines the problem and its inherent research questions, which are to be answered.

On the topic of Text Mining, a search through the current literature yield limited results in regard to the amount of relevant research articles that show practical uses of text mining in business environments. However, the research does present TM as being of business value since the research can be used to derive implicit evidence to such claims (as presented in Section 2.2 and its subsections).

Proving the business value of TM is the fundamental problem of this study. The findings and their validity to give claim to the business value of TM, are dependent upon how well an accurate and suitable interpretation of such claims is made possible. There are no indications for when business value is derived, or how it is decided upon, other than what can be concluded by the implications of the research. To this end it is possible to create a concept matrix, where the concept of the term is defined and captured, capable of acting as a model that could decide when business value is derived from TM. This proposed concept matrix was earlier presented by the author (Section 2.2.).

As identified, there are several possible business domains and data sources to extract information of business value from, and the literature demonstrates TM technology as being of value. However, no research has had the purpose of proving such claims by investigation. There is an apparent gap in the research, where it would be of interest, help, and arguably necessary for businesses to know what specific business value can be derived from their available Dataset, before implementation of a technology such as TM is ensued. This would entail the additional need of addressing, and considering the problem of data sources, and business domains not guaranteed being similar. It is thus a requirement to investigate businesses having similar types of Dataset, and by identifying general business areas where TM can provide business value. This could potentially give claim to the business value of TM across different business domains and data sources, with the prerequisite of having access to a Dataset of a format which is general and suitable for extrapolation.

3.1. Problem statement

The section present the problem statement and its inherent research questions. Each research question is given explanation to their purpose and importance.

- What is the business value of Text Mining?

The concept matrix is used to capture the concept of “The Business Value of Text Mining”, to solve the fundamental problem of the study, proving the business value of TM. The reason is that there are no other suitable course of action to decide when business value is derived. The concept is grounded on how business value is emerged.
throughout the literature; and as such the literature forms the basis for when business value is derived. Thus the following questions are possible:

- What are the general business areas where TM could be applied to derive business value from?
  - The question entail and require the investigation into businesses, to identify potential, and general business areas, where TM could be applied to derive business value from.

- What does TM analyses reveal about the investigated Dataset, and what does the analyses say about the business value of TM, given how well they agree with the concept?
  - To clarify, the question encompass the use of TM for the extraction and presentation, of information of business value, on the given Dataset. The purpose is demonstrating and proving the business value, of this sample, given agreeable results from analyses of Dataset with the business value categories of TM, i.e. CR, QC & QA, and CST.

3.2. Limitations of the study.

There will be no creation of algorithms, therefore, differentiating from much of the research content already conducted and largely available in the area of TM. In the context and with the intention of proving the business value of TM, it is not appropriate to create an algorithm to quantify the precision of analyses on data. Creating an algorithm would be more appropriate post hoc, as the findings might give implications on how the derived of business value from TM, could be improved, or why not according to expectations, to make way for writing a more suitable algorithm for business purposes.

There will be no integration and innovation of software. All software which is written, is by ready-made packages and libraries, to strictly conduct the necessary analyses for proving the business value of TM. The chosen programming language R, has a high variety of ready-made packages and libraries suitable for the technology of TM. Possible limitations of the R-Language, are inherited. Resources are a factor to limitations of the study, therefore, there will be no use of TM tools which requires paying for a subscription, or are not Open-Source. The study is carried out in a limited time frame, writing fully customized software, for deployment, would bring about unnecessary time before enabling the investigation of the specified research questions, writing the software could possibly (author note: allegedly) be a case specific problem, to study, in of itself.

Competency is an additional limitation to the study. Findings are dependent upon how competent the one conducting TM is with the technology.
Adhering to the limited time frame, there will be no interviews, beyond the contact with the companies involved. Collection of data is done through a SMS-survey, considered suitable to answer the research questions, to keep within the limited time frame. Further limitations are the inherent limitations following the use of the SMS-survey for collecting data. Time also limits the investigation into the possible available types of Dataset, the different business domains, and data sources that exists (i.e. it will not be possible to cover every aspect of TM in a business setting).

Collection of text documents, i.e. the Dataset for analyses with TM, are limited by the approval and access to said Dataset, by the companies where such data is extracted from. The limitation for the fact that companies do not share sensitive, and possibly competitive data so freely. This implies smaller sample Dataset available for conducting the analyses on.

3.3. Expected outcome

The main expected outcome is giving a suitable answer to the main question simultaneously solving the fundamental problem. The latter by the outcome of proving that business value can be derived, by the use of TM. Further outcomes are expected to be novel patterns and meaningful information from TM. An outcome of outmost importance is the discovery of possible or necessary actions to the improvement of a service or product, in the given Dataset. Followed by the presentation of said discovery. This would imply that business value is gained for a third party, and the general research question can be given evidence.

By the means of a SMS-survey, the expected outcome is the identification of business areas where TM could be applied to derive business value from. Therefore, proving general business value for other parties as that which is true for an analyzed Dataset, is also true for other parties in the general business domain.
4. Method

In the context of a research project Berndtsson, et al., (2008, p. 12) define method as the following: “a method refers to an organized approach to problem-solving that includes (1) collecting data, (2) formulating a hypothesis or proposition, (3) testing the hypothesis, (4) interpreting results, and (5) stating conclusions that can later be evaluated independently by others.”

The following chapter has the intention of outlining the chosen research method, and its inherent means. Therefore, answers shall be given to why literature is selected; furthermore, what other forms of data is collected, and description of said data, followed by establishing their purpose and how they are to be used. It is followed by description of how analyses are to be conducted and also the adopted research ethics.

4.1. Text mining case study on dataset of survey feedback

Bhattacherje (2012, p.93) defined a case study as: “a method of intensively studying a phenomenon over time within its natural setting in one or a few sites. Berndtsson, et al., (2008, p. 62) complements the latter with: “especially suitable when there is a desire to understand and explain a phenomenon in a field which is not yet well understood.” It has several methods to data collection, and inferences about the phenomenon of interest tend to be rich, detailed, and contextualized (Bhattacherje 2012, p. 93). The phenomenon of study in this case study is the: (1) use of TM technologies in a business setting, (2) to derive information of business value.

This case study employ data collection methods such as: secondary data (data collected for other purposes) for drawing inferences with TM technologies. This case study is employed in an interpretive manner for theory building (Bhattacherje, 2012, p.93), and argues for its theory building as there are no prior similar theories identified, to the knowledge of the author.

Case studies have their inherent weaknesses (Bhattacherje, 2012, p.93). As Bhattacherje (2012, p.93), the author predict these to be due to heavily contextualized inferences. The latter because the secondary data used in the TM, and the SMS-survey, could demonstrate business value for the current organization and context, yet show difficulties in generalizing inferences to other contexts or other organizations. However, this could be established with corroborative case studies (Bhattacherje, 2012, p.101). An additional weakness is the replicability of the results. Considering the Dataset, to replicate the TM analysis results might indicate difficulties in observing the same phenomenon, given the uniqueness and idiosyncrasy of the given case site (Bhattacherje, 2012, p.101); however, the conclusions of the case research may be possible to replicate. Induction is used to learn from the collected data and build the concepts.

This study lean more toward gaining understanding over the phenomenon of interest, therefore one could say it is more of an interpretive approach to the collected data.
Vidgen & Braa (1997) characterize the view of interpretivism as being “concerned with making a reading of history in order to gain understanding”. According to Vidgen & Braa (1997) in interpretive approaches the researcher attempt a minimal impact on the situation, reducing possible interventions/change.

By conducting TM analyses with the use of the R-Language, the investigation has the objective of extracting information of business value from the Dataset, such as the problematized areas on different features or aspects of the given courses. The latter to establish the second dimension, capable of providing evidence in proving the business value of TM. Thereby, the derived information are expected to be means of quality improvement and increased customer satisfaction, for the given service; additionally, a suitable presentation of the information is also of importance. The results of the text mining has the goal of extracting information (of business value) belonging to any of the three TM categories of business value (i.e. CR, QC & QA, and CST).

Its related research question is: “What does TM analyses reveal about the investigated Dataset, and what does the analyses say about the business value of TM, given how well they agree with the concept?”

4.2. Participants

**IP.1**

The primary collection of data are all conducted through the company IP.1. Networks AB (IP.1). IP.1 is the service provider of a Business Intelligence-tool called “AnalysSMS”, giving its subscribers the capability of sending digital surveys, via SMS, thereafter, being answered with a smartphone. The surveys are fully customizable, capable of asking any type of question, through customization, as well as mixing these different types. This can result in either quantifiable questionnaires with ratings and selections, to semi-structured questionnaires with freeform comments and answers, to being fully freeform not specifying any answer.

Further, the data collected from the participants are real-word, and relative to the context, therefore, supports in adding the desiderata of realism to the study.

**Traffic School**

IP.1 who has, as of yet, not implemented a way of automating an analysis of the freeform comment-fields, state that it would be of high interest to their customers in having an integrated text analytics tool for such a purpose. The case study is focused on investigating one of their customers, a traffic school who collect data from their students using surveys after the students have gone through a specific course. Similarly to the queried subscribers of “AnalysSMS”, the traffic school also employ this tool for survey data, therefore these sources become highly relevant to each other.
4.3. Mini-literature review

By conducting a mini-literature review, which earlier was presented, the first dimension of the study is established. It has the underlying purpose, of supplying the study with a theoretical anchor, and simultaneously delineates the domain. It seeks to verify that the chosen main topic, which is the business value of text mining, does not exist in the literature, therefore, give backing to, and support of, an execution on the succeeding objectives.

In the literature selection, many keywords for article searches were intuitive and top-of-mind, by making considerations to the current context. The resulting keyword in searching for articles are some of the following: “opinion mining and sentiment analysis”, “text mining case study”, “text mining techniques”, “business value text mining”, “text mining defect”, “opinion mining and sentiment analysis”, “summarizing text mining”, “text mining big data”. All selected from sources such as Springer, ScienceDirect, ACM, Google Scholar, nowpublishers, WorldCat.

The decision on which articles are of relevance, is based on how closely related they were to the topic, and if selected research had any prior reference to the same articles. Some of the articles were also selected, and discovered, by following the trail set out by other selected articles, resembling a snowball effect. The latter if one article mentioned something of higher relevance to the subject, or generally was a stronger source to explain concepts. The snowball effect also aid in establishing and delineate the current research domain. Some articles were also recommended by word-of-mouth.

4.4. SMS-survey method

An empirical investigation of the subscribers is conducted with the use of a SMS-survey method. It is a standardized questionnaire to collect data about thoughts and behaviors. According to Berndtsson, et al., (2008, p. 63) a survey is suitable if you want to explore perceptions concerning a specific, well-known methodology. The SMS-survey for example attempts two things, (1) identify more categories of business value for TM, (2) the investigation of the methodology regarding deriving value from their information derived by using AnalySMS.

To clarify, the SMS-survey has the purpose of establishing a third dimension, identifying areas (i.e. categories) in various businesses, to find out where business value from TM could be derived (To avoid confusion, this data is not analyzed with the use of Text Mining). Assumptions are made such that, if the companies collect suitable information via their freeform comment fields, TM could be applied to derive business value from such a source. It adds to the empiricism by finding new business areas where TM could be applied, insofar as the results of the SMS-survey makes it possible. It requires the TM case study to be conducted beforehand, since insights gained from the latter are extrapolated, to infer the applicability of the technology onto the described data.
Its related research question is: “What are the general business areas where TM could be applied to derive business value from?”

4.5. Analysis

According to Berndtsson, et al., (2008, p. 83), the most systematic way of analyzing the collected data is to go through each of the objectives at a time, evaluating said data against responding objective. The proposed research questions are conveniently divisible by their framing, and data are collected with the specific purpose of enabling appropriate answers to each of the said research questions when the goal of the objectives are accomplished.

The analysis is conducted in accordance to a predetermined sequence. (1) First a review of the literature, (2) a TM case study is conducted, (3) an SMS-survey is conducted. Each dimension and their initial conclusions, i.e. the results, are compared and conjoined for drawing the inference, allowing for the argument to the evident business value of TM. An inductive approach is used on the data to make the conclusions. Bhattacherje (2012, p. 15) define induction as “the process of drawing conclusions based on facts or observed evidence”. The intention is to improve the credibility and confidence of the study, by demonstrating triangulation across the collected data (Bhattacherje 2012, p. 110).

A model is used to clearly illustrate how the analysis, as a whole, shall be conducted (Observe Figure 3.). (1) By using the data with its associated objective, answer each of the initial research questions, (2) derive a result as emerged from the answers, (3) compare the results to answer the main questions, (4) a solution to the fundamental problem is emerged. With the steps (3) and (4), the aim of the study is considered to be achieved.
4.6. Research Ethics

The study adopts research ethics because: “science has often been manipulated in unethical ways by people and organizations to advance their private agenda and engaging in activities that are contrary to the norms of scientific conduct” according to Bhattacherje (2012, p. 137).

4.6.1. Anonymity and Confidentiality

Excluding IP.1, it is per request, that all data gathered or supplied from the third parties (subscribers and traffic school) are ensured confidentiality and anonymity. In accordance with Bhattacherje (2012, p. 138), “To protect subjects’ interests and future well-being, their identity must be protected in a scientific study”. Actions are made for anonymity and confidentiality, to ensure that no specific individual is tracked or profiled. Involved companies and names are likewise given the same considerations.
4.6.2. Disclosure

The subjects’ for collection of the Dataset, and SMS-survey data are provided information about the study, its expected outcomes, and the potential benefits of the results (Bhattacherje 2012, p. 139). Furthermore, the subjects’ are asked for a decision of participation. Supplying information on the topic is necessary as the field is seemingly new. Disclosure of said information has the motive of adding to the subjects’ knowledge-base, as a complement, in giving better answers for the SMS-survey. The study consider the potential biases in subjects’ responses, in the SMS-survey (Bhattacherje 2012, p. 139).
5. Research execution

The chapter give a more detailed view on the information concerning the execution of the research method presented in the prior section. The Data source selection and choice of data is a crucial element to conducting TM analyses, and as such is qualified to its own subsection. Subsections 5.2 and 5.3 give detailed descriptions of techniques and methods for as they are employed in TM phases. 5.4 give commentary to decisions regarding the design of the SMS-survey and reports how the collection went. The mini-literature review is presented in section 2.2 for convenience, and should be digested in this manner.

5.1. Mini-literature review

By a mini-literature review, a systematic identification, selection and extraction of business value from the literature, followed by, categorization of each of the different “business values”. Categories of business values are formed with an inductive approach. This in order to make it easier to grasp the concept of The Business Value of Text Mining. With the concept matrix, the aim is to further clarify the concept (Bhattacherje 2012, p. 21). (1) The content is generalized to fit the context of business value for TM, (2) Divide and choose, by increasing the granularity, thereby, identification, selection and extraction of business value from each article, for contrast and comparisons, (3) Decide on the suitable category of business value, (4) Sort each article in the according category by: associated benefit, insight & knowledge, and origin of its related domain/setting and data source.

5.2. TM Data source selection

In their five-stage method for text analysis, Rose & Lennerholt (2017) place emphasis on the phase for selection of data sources, with the following argument: “In research on developing and testing new algorithmic techniques in the text analytics field the choice of data source may be relatively insignificant, however this is not the case in research in other fields where text mining is used as the research method.” Rose & Lennerholt (2017) give further weight on the issue of studies in text analytics, where there are assumptions that the chosen sources represent “what happens on the net, and that what happens on the net also represents the physical world”. They therefore stress the importance of the data sources to: “ideally be representative and relevant”, as the results of the analyses are affected by the composition of the sample text sources. In their conclusion, Rose & Lennerholt (2017) give comment that the collection of data from the net, should be “in sufficient quantity to make it impractical to use conventional qualitative analysis techniques in response to an open research about trend in business intelligence.” The problem of sampling is automatically addressed by a larger quantity of data. Here some clarity is gained regarding the difficulties, and lack of evidence to the business value of text mining in the literature, supporting preconceived notions that much of the research prove TM to be of value, however, research show lack of evidence to give claim to such value in a business
context (cf. section 3). The latter for the reason that the representation and relevance of the data not being addressed to prove the business value, with the exception of He, Zha & Li (2013).

It is possible to establish a criterion for selection of data source to address the principle issues identified by Rose & Lennerholt (2017), and to ensure validity to the results of the analyses with the following:

- Relevant to current subject
- Represent the physical world
- Data quantity

5.2.1. Choice of Data

The Dataset are collected from a database of the company IP.1, containing data necessary for the study. The owners of the data are asked for consent beforehand; names in the data are removed to not show up in analyses; translations are made for the excerpts, since the data are in Swedish, and to further the anonymity (the translations do not affect the semantics, given the authors native tongue is Swedish). The dataset gathered consist of survey feedback from freeform comments. The dataset from the traffic school has three standardized questionnaires; one of these differs in its format, this does not affect the textual data. The questionnaire address various features of the courses which are of interest for the traffic school. The latter is numerical data, viable for statistical analysis. The earliest collected data in the dataset are from April 2016, and oldest entries being February 2017. The data rows, quantitative questions included, do amount to well over 200,000; thus unique entries per respondents and sum thereof are too numerous to pinpoint exactly, and effecting such a task would be redundant, and time is better spent on the actual textual analyses.

Key-features of the questionnaire are numerical data from a scale of 1-100 on: how good a teacher was, how good is the room/locale, how good the customer service was, how good was the premium package, how good was a particular parts of the courses, information regarding payment/booking/costs etc. It is possible to draw some conclusions upon the information, however, one apparent weakness is the lack of nuance to the specifics, in regards to why a student gave a particular score on each feature. For example, to know that the customer service has been given a lower score of 30 (scale 1-100) does not give specifics on “how” or “why” it could/should be improved; instead it can only give the answer: the customer service (a feature) should improve (is less than optimal).

The abovementioned therefore highlights one of the strengths of the dataset because of its additional option, a freeform comment field for general feedback, which inquire the students for further qualitative feedback. The comment field supplements the
surveys, as it enables its respondents to give nuanced qualitative insights to the features which are addressed, or not addressed by the numerical data.

Considering the context of this study where the aim is to investigate and evaluate the business value of TM, the data source is in respect to the abovementioned criterion, and to the notion of information of business value, highly representative and relevant to such an objective. The author judge the representation by it being collected from actual customers, also being used by the traffic school who use the data to better their own services, adding to the desiderata of realism to the data source. The relevance to the current subject is judged by it being relevant in a business context, where data were collected with the intention of gathering information to derive business value from it. Considering the data quantity, the author judge it sufficient due to it being highly impractical using a hermeneutic style to qualitative analysis (establish a relationship between the whole text and its parts) since, similarly Rose & Lennerholt (2017), the Dataset were too large to develop some resemblance of coherence over the content.

5.3. Text Mining the Dataset

Before beginning the TM analyses the Dataset go through the process of cleaning and subsetting the textual data from that which is irrelevant. Prior to preprocessing, the Dataset follow the same format, having a total of 11 columns such as date, entry ID, questionID and more. For the creation of a TM corpus these are viewed as noise in the data and must be scrapped for such a purpose. The data are imported to the programming language R, where irrelevant columns are dropped. Remaining is the single column called “Comment”, containing comments. This step is reiterated for the total four different Dataset which are chosen for the TM analyses.

Using the R package called tm (Feinerer, Hornik & Meyer., 2008), the Dataset are combined and converted into a text corpus. Having created a text corpus, further cleaning of the remaining column is conducted. NULL values, punctuations, numbers, special characters, whitespace and swedish stopwords are removed. The remaining entirety and all rows are also converted to lower case characters. No word stemming is conducted as it had no impact on found patterns. Considerations are made to some of the entries being misspelled and therefore affect the ensuing analyses. The text corpus is treated by the chosen TM methods as a whole and does not aim to find differences or compare each document. The entries are all authored by Swedish speaking students, and therefore the results are all in said language.

The results of the preprocessing step is a text corpus with a total of 4688 entries (rows in the data).

Following the creation of a text corpus, a Document-Term Matrix (DTM) is created from it. The Document-Term Matrix is a structured data format to represent texts for computation (Feinerer, Hornik & Meyer., 2008), necessary for the later analyses. The DTM could be described as having each document on a row, and terms as columns.
The method for the TM analyses, in this study, treat the DTM as a bag-of-words, meaning the words are viewed as a bunch of attributes of the documents, with no attempt at semantic parsing.

The conducted analyses split each of the character vectors in the DTM tokenized string. Unigrams, bigrams or trigrams are created, depending on what has been specified in the N-gram tokenizer from the Weka toolkit in R, RWeka. For example a unigram in this sentence would be distinct words (i.e. grams) such as “for, example, a, unigram”, while a bigram would be the probability of the grams co-occurrences “for example, this sentence”, where the former and the latter have a higher probability of occurring together in a sentence. Unigrams are not expected to reveal any key-insights, however, they fill the purpose of giving an overall impression of the documents and identify key words which might be irrelevant. Similarly to Rose & Lennnerholt (2017), iterative passes with the data were made to remove additional irrelevant words, these were unadressed by the Swedish stopwords list in the tm package. Particular words such as: education, drivinglicense (no direct translation from swedish), traffic, course (the course from swedish), course, more, back, think, it, manner, half, alot (swedish has abbreviation for the word alot), walk, got, three, else, will, take, alcohol, drugs, whole, men, women, thought, drive, driving, show, very, things, student, held, should, were removed. These words were removed either by being revealed in the pre-analyses or by being judged as irrelevant. To clarify the latter with an example, the word “drugs” is removable for the reason that a bigram such as “drugs bad” would have no impact considering the context. The latter could however indicate whether the students got the intended message of the course.

As per its definition, TM attempts finding meaningful patterns in textual data. Therefore the study does not find the less frequent terms of the DTM to be of interest. The terms which are the most frequent are viewed as being meaningful patterns, and the initial analyses therefore discards the less frequent ones.

5.3.1. Analysis of TM results

Initial analyses are conducted by making interpretations of having visualized the manipulated data using graphs for wordclouds, and graphical barplots. With the now at a glance understanding of the data, relationships between the different patterns in the entire DTM is found.

Correlation analysis is conducted on the identified patterns. Correlation analysis is used to validate prior patterns and interpretations through validation using an additional TM-technique. The correlation analysis can give support to interpretations and identified patterns, through some sort of pseudo-triangulation method for TM. It would also work as an independent technique to derive insights from the data.

To validate the initial interpretations, and preconceived notions, or to gain further insight on the meaning of a specific pattern, backtracking is made by search through the raw data. The latter is done by having the qualitative analyst manually read
entries that are related to a specific pattern. An additional check is made for the actual numerical correlation between each term.

**Determining the business value of the derived information**

To know when derived information is of business value, deductive reasoning or inductive reasoning is used to infer the nature of the information; that is, compare and see if it matches any of the categories of the concept to make the inference that it is of business value. The premises for the argument are “derived information matching the categories of BVTM”, and “information is of business value”; where the proposition is that the former would infer the latter.

The deductive argument would have the conclusion: “derived information match the categories, so we can be certain the information is of business value”.

The inductive argument would have the conclusion: “derived information match the categories, so it is probably information of business value”.

The author view the propositions as equally valid, and both are suitable to decide business value for each of the insights that were derived. To the conclusion of the derived information being of practical business value, it would entail a positivistic approach; thereby taking action upon the information that is derived, and measuring the impact of said action. The author argues the latter not being necessary, since the concept can decide whether information is of business value more generally, thus possible to extrapolate the evidence to some extent.

To clarify, the author make the inference that, given the derived information from TM match any of the categories (i.e. if it is true), then it is information of business value. The figure reads as a conditional statement (Observe Figure 2.). (1) TM is conducted, (2) Information is derived, (3) with the help of the concept see if it has a matching category, (4) if the latter is true, make the conclusion that it is Information of Business Value.

---

Figure 3. Illustration of how the inference to information being of business value is made.
5.4. SMS-survey

The SMS-survey is designed to query the various subscribers of the software tool AnalysSMS. These subscribers are companies, and therefore complicated the end expected response frequency, and forced the need to make consideration whether a lengthy questionnaire is suitable, given this particular type of sample. The questionnaire is also written in Swedish.

The questionnaire response format is designed to capture nominal responses, more than two unordered options (Bhattacherje 2012, p. 75), and aggregate the nominal responses with open-ended questions, that build upon the prior queries. Considerations are made to the sequencing of the questions, and aim to achieve some priming by its logical flow. Initial questions are factual, to place a certain type of object in their thoughts, followed by more specific questions to gain specific information about the object in mind. Nominal questions divide the categories found in the concept matrix to make precautions to the different types of data that the respondents might be thinking of.

The survey follows the recommendation by Bhattacherje (2012, p. 75), that the question content and wording is sensitive to the types of questions asked, and each to be carefully scrutinized on a set of given issues. To follow this ruleset by Bhattacherje (2012, p. 75) is the ideal, the author argues for some of its deviations. Neutral wording by avoiding positive, and negative orientation. Using simple language by minimizing the use of technical jargon. Reducing question ambiguity by supporting descriptions of what is currently being referred to. Avoid questioning bias, as TM and its value is not mentioned. Avoid double-barreled questions by asking questions with single answers instead. Not being too general, and instead be specific on what type of information is being asked for. Reduce question detail by having removed unnecessary detailed questions. Presumptuous questions are hard to avoid, the survey contain one presumptuous question, asking for the limitations of manual analysis. Avoid imaginary questions, no question is imaginary. Necessary information is supplied beforehand, to specify what is asked for and gain some accuracy in the responses.

The questionnaire was sent to the respondents by SMS, possible with AnalysSMS:

1. Identify the type of information that is gained from the respondents.
   
   “What type of information do you gain from the collected freetext responses?”

2. Specify if this information matches any of the categories of business value of TM or discover additional categories.
   
   ”Does this type of information fall within any general business domains?”
3. To gain understanding of the information they have in mind, by query of what type of insights are gained from it (i.e. it has the intention of making it clear why the information belongs to a certain category).

“What type of insight is gained by the information you are thinking of?”

4. Limitations to their current way of analyzing the data.

“What are the limitations to how you currently derive insights from this information?”

5.4.1. SMS-survey analysis

The SMS-survey was sent to a total of 38 respondents, whereas a total of 4 had given an answer. 3 opened the SMS-survey and did not answer. Given the focus of the survey to collect qualitative data, approximately 40 respondents seemed reasonable, given the assumption to receive at least 10 responses. The respondents of the SMS-survey were selected by IP, the set prerequisites was the respondents being users of the tool AnalysSMS. Looking at the respondents who received the SMS-survey (sensitive information that is not shared for reasons in regards to ethics) the given sample differentiates in terms of type of domains (the business domain in which the respondents belong to). The SMS-survey lets the respondents select multiple choices on the multi-optional questions (Q1 and Q2). To give mention if they have several different datasets. Furthermore, the respondents were given the option to not include their respective domain, if they find this information sensitive.

Analysis of the SMS-survey was made with an inductive approach. The analysis of the SMS-survey results were conducted after the TM case study. Considering the low response, it was not necessary to use any advanced qualitative techniques for the survey analysis. The numerical data was automatically visualized by the survey-tool, in the form of bar graphs, with the intention of finding possibly new categories or business areas where TM could possibly provide business value or support existing categories from the concept matrix. The qualitative answers were analyzed by basic reading and then made comparisons using the understanding gained from TM, to see if the described type of data could be viable for TM. This sort of analysis is possible since the data format for the type of data being discussed in the SMS-survey has similarities to the traffic school Dataset. The latter, considering the participants in the SMS-survey and the traffic school both employ the tool “AnalysSMS”.
6. Analysis

The chapter presents the analysis of the conducted investigation, into the business value of TM.

6.1. Text Mining

The subsection is divided into different subsets readable as steps in the method. Each could be viewed as different paradigms to deriving business value with the use of TM. The wordcloud are originally in Swedish, the author translated each wordcloud, manually, into English.

6.1.1. Finding patterns by N-grams

![Figure 4. Bigram of the DTM.](image)

The wordcloud for bigrams, (Observe Figure 4.) reveal several unique patterns. Some bigrams are reoccurring in a different wording, with similar semantic meaning. Identified patterns are:
• Long time (too much time/time consuming) - 84 Occurrences
• difficult hearing – 17 Occurrences
• bad air – 17 Occurrences
• hot/warm room – 14 Occurrences
• bad ventilation – 11 Occurrences
• cramped room – 11 Occurrences
• cold room – 11 Occurrences

The pattern “long time” is being referred to several times in different bigrams, in different wordings, “long time” is chosen as it is the most frequent.

The other patterns are judged as being of little value in terms of prospect for improvements to the courses, or having low frequencies in the data, they are ignored for now.

Figure 5. Trigrams from the DTM.

The wordcloud for trigrams help strengthen the validity of the patterns “long time “poor air”. The trigram patterns do however not reveal anything new or give any further insight.
A final N-gram is created, a quadgram, and a new pattern is emergent. Similarly to the trigram much us repeated, except don’t drink and drive and “pain back neck chairs”. The former pattern says nothing new other than the students’ aversion to the mention of the phrase “don’t drink and drive” starting to get old. The other pattern “pain back neck chairs” is however relevant.

### 6.1.2. Correlation analysis of found patterns

Several “themes” or “patterns” are emergent by the prior analyses. The wordclouds for the bigrams followed by the trigrams, and quadgrams, provide at a glance understanding, where preconceived notions about the data are formed. However, the author considers the potential biases and having an instinctive lack of trust for the data force the need of further inquiry.

The numerical representation of these “themes” and their correlation to one another, is found using the function cor, to get the correlations in a matrix, in R. Using the correlation function, it is possible to specify any term and compare its correlation to any given term. The document is stemmed, followed by collapsing similar terms.

The matrix of correlations looks like this:

<table>
<thead>
<tr>
<th></th>
<th>cramped</th>
<th>enjoy</th>
<th>warm</th>
<th>cold</th>
<th>poor</th>
<th>difficult</th>
<th>long</th>
</tr>
</thead>
<tbody>
<tr>
<td>room</td>
<td>0.29706</td>
<td>0.04274</td>
<td>0.35172</td>
<td>0.15819</td>
<td>0.19064</td>
<td>0.04501</td>
<td>-0.30982</td>
</tr>
<tr>
<td>time</td>
<td>0.01656</td>
<td>-0.00911</td>
<td>-0.00374</td>
<td>-0.02255</td>
<td>0.00371</td>
<td>0.01391</td>
<td>0.27805</td>
</tr>
<tr>
<td>air</td>
<td>0.04453</td>
<td>-0.00352</td>
<td>0.08668</td>
<td>-0.00623</td>
<td>0.26044</td>
<td>0.05225</td>
<td>-0.00758</td>
</tr>
<tr>
<td>hearing</td>
<td>-0.00961</td>
<td>-0.00468</td>
<td>0.00902</td>
<td>0.00077</td>
<td>0.03276</td>
<td>0.29560</td>
<td>-0.00643</td>
</tr>
<tr>
<td>ventilation</td>
<td>0.02339</td>
<td>-0.00500</td>
<td>0.12903</td>
<td>0.03202</td>
<td>0.22197</td>
<td>0.02738</td>
<td>-0.00321</td>
</tr>
</tbody>
</table>

Figure 7. Correlation for various terms.
The correlation between the terms “long” and “time” is first investigated, resulting in a positive correlation of 0.278050229. It is possible to claim the prior pattern to be valid and in accordance with the preconceived notions.

**Room**

The correlation analysis for the term ”room”, show a high positive correlation to the terms: “cramped”, “warm”, “cold”, “poor”. Terms “room” and “warm” show a high correlation of 0.351720699, which in comparison to “cold” showing only 0.1518195859.

Therefore, this analysis give implications that the heating is the dominant issues to how well a students’ experience is for the courses. However, both are valid for further investigation.

Additionally room show a correlation of 0.29783683 to the term “cramped”. 0.29783683.

**Air & Ventilation**

The terms “air” and “poor” show a correlation of 0.265445616. The term “ventilation” and the “poor” show a correlation of 0.221376708. Terms “ventilation” and “warm” show a high correlation of 0.129034764.

The items in question refer to the same thing. Ventilation is poor/bad, therefore air is also bad, ventilation is bad therefore it is hot.

**Hearing**

The terms “hearing” and “difficult” show a correlation of 0. 0.29560335.

Initial conclusion is that the students experience difficulties in hearing what the teacher is saying.

**Pain back neck chairs**

The pattern refers to pain in the back and neck, referring to the uncomfortable chairs. No correlation analysis should be needed, as the author find the N-gram quite self-explanatory; however, following up on it show the term “pain” to be highly related to similar words “neck” 0.37, ”butt” 0.35, ”back” 0.29, “chairs” 0.28.

6.1.3. Validate findings using the raw data.

The correlation analyses prove fruitful, and do show some validity for the initial conclusions drawn of the data. A relation between the terms exist, both in the bigrams and trigrams, similarly showing some consistency in the correlation analyses. However, only knowing what patterns exists and drawing shallow conclusions upon them, would not be enough to be certain on why the students give certain types of feedback or how improvements can be made. To validate the patterns
and derive actual insights from them, should entail extracting some nuance of the data, which is possible with supplementary analysis using simpler qualitative techniques. To validating the findings in this manner, for actual insights, is done by manually reading the related entries. High level qualitative analysis techniques are not required, basic sense-making is used to get the semantics of the entries. The excerpts from the raw data are translate to English.

**Time**

Searching for only “time” would give too many hits, the search is conducted with the bigram “long time”. The findings reveal the related entries referring to the students’s being generally dissatisfied with the length of a course. An excerpt from the raw data would clarify:

“Time consuming, the same content could have been digested on a much shorter time. But it might be something not under the influence of the educator.”

“Time, 3h is too much for the information that is necessary. Shorten the time and skip the pause.”

By search of, and by manually reading, the related entries in the raw data it is apparent that the entries are, with negative orientation, referring to the length of the course.

**Room**

By further analysis, going deeper by searching the raw data, we divide the months into “warm” and “cold” to validate the findings:

33 “warm” summer.

9 “cold” summer.

15 “warm” winter.

15 “cold” winter.

The results show that “warm” is a frequent complaint, occuring closer to summertime, while “cold” more frequent closer to winter. The conclusion of the analysis is that regulating the heat is an issue that should be considered for the betterment of the students’ experience. Aggregating the data with the actual temperatures during these dates, would supplement and give a more precise explanation to the why these findings appear. A weakness in the conclusion is that the analyst base seasons according to months and guessing whether the weather is hot or cold, the external factors based upon guesses. Excerpts for hot and cold:

“It was too hot, which affected the lecture.”
“The teacher was calm and good to listen to, the room was very cold so you were freezing.”

Figure 8. Aggregate three out of four matching Dataset from numerical data.

The figure above reveal Q6 (question 6), in the numerical data having the lowest score. Q6 relates to the question of what the opinions of the room is. It is capable of telling the analyst that the room is bad, however, it fails in giving information regarding why it has such a low score and how it could be improved. In this respect TM is the better option.

**Air & Ventilation**

The analysis concludes the air conditioning to be a factor to why numerous complaints are received regarding the heating in the room. Excerpts are:

“The only thing a bit worse was the ventilation in the room. It got very warm and stuffy with so many in the same room.”
“Bad ventilation in the room. Warm and stuffy.”

"Warm and bad air in the room."

One thing that is noticeable by searching the raw data, is a weakness to the way the patterns are identified. There are additional patterns indicating the bad ventilation or poor/bad air, which are found by searching for the term “air” for example:

“Air conditioning would be great...”

The excerpt has a negative orientation however, since the technique used for TM form patterns based on words being similar, not by having words with same semantic meaning, therefore, this relationship is not found. Advanced NLP-techniques for sentiment mining, would probably find a more frequent pattern to the relationship of “luft” and “ventilation”.

**Hearing**

The pattern “difficulties hearing” is the harder pattern to decide upon what the dominant cause to said difficulties. Searching the raw data point toward different causes:

“The teacher was good but at times he spoke too quietly so it was hard to hear. “

“The speakers were very bad. Difficult to hear what was being said on the film. Otherwise everything was good!”

“Difficulties in hearing as there were people who were interpreting loudly to those who did not know Swedish”

The above citations are excerpts found in the raw data. It is possible to identify what the causes are, however, there is too much variability in the data to be precise in in knowing which cause is the most problematic. The preconceived notion of the issue surrounding difficulties in hearing is here proven incorrect as there are several causes to this pattern, that is, it is a cause not the cause.

**Pain back neck chairs**

Searching for the raw data is done with the keyword “chairs”. The results show various entries not all using the same wording of “pain” or “back”, “neck”, however, the entries refer to the same issue. The following excerpts for example:

“It was very educational but the chairs were pretty hard.”

“The course chairs were not comfortable. Other than that everything about the room was good.”

“Get pains in the back and neck by the chairs in the classrooms”
“Tiny room & very long meeting, so the chairs should've been softer... Believe me, you get some real pain after those three hours on the hard chairs”

The results show the chairs to be an issue needing to be fixed. The initial pattern does not encompass every possible sentence in combination, but did succeed in help discovering an issue to be addressed.

6.1.4. Sentiment Mining

Sentiment mining was conducted on the Dataset, to see if any additional insight could be derived, with the help of the tidytext package. However the results were either equivalent to prior methods or less fruitful. A custom sentiment lexicon for the Swedish language was created from a sentiment lexicon available on the web. The contextual issues for semantic lexicons are apparent, e.g. long is assumed being a positive term, when in fact, in this context it is actually negatively oriented. This could be fixed by the one doing the TM, though the issue of time being a limitation to the study kept this from being addressed. Therefore, given that one would have a more exact sentiment lexicon, it would be viable to follow up on a term, such as “interesting” (Observe Figure 9.), and find its correlations to other words. This would mean a query as such as: What is interesting? And as such the sentiment mining results could be beneficial to the analyses.

The encoding of the Dataset characters brought up some issues when creating bigrams for the sentiment mining, creating issues such as “ÄÄÖ” ending up as weird symbols. The preferable character encoding for creating bigrams with tidytext was “latin1”, whereas the imported Dataset were encoded as “unknown”.

The author also attempted translating all of the textual data with the help of the translate package through the “Google Translate API 2.0”, and employed the same sentiment mining techniques. Here the character translations returned better encoded data, also suitable for bigrams; however, it is uncertain how trustworthy the translations are, and the author concludes the findings as fruitless (Observe Figure 9, Figure 10, Figure 11).
Figure 9. Sentiment mining for positive unigrams on textual data translated to English.
Figure 10. Sentiment mining for negative unigrams on textual data translated to English.
Figure 11. Graphical plot of negative and positive sentiment with \textit{ggplot2} package. They are translated to English.
6.2. SMS-survey

6.2.1. Nominal responses

| Feedback/opinions on product | 08/05/2017 R1 |
| Feedback/opinions on company | 09/05/2017 R2 |
| Feedback/opinions on product | 09/05/2017 R3 |
| Customer needs | 09/05/2017 R4 |

Figure 13. Question (1): What type of information do you gain from the collected freetext responses? Identify the type of information that is gained from their respondents (R1-R4 implies which respondent).

| Customer Relationship | 08/05/2017 R1 |
| Customer Relationship | 09/05/2017 R2 |
| Quality Control | 09/05/2017 R3 |
| Customer Service | 09/05/2017 R4 |

Figure 14. Question (2): Does this type of information fall within any general business domains? Specify if this information matches any of the categories of business value of TM or discover additional categories.

By the results of the nominal data, it is easy to conclude that the subscribers use their survey-data results in areas such as Customer Relationship, Customer Service, Quality Control and Quality Assurance. It is clear that the categories of the concept matrix can be improved, since the nominal data give implications to there being an
obvious distinction between Customer Service and Customer Relationship. Quality Control and Quality Assurance should similarly be separated as they too are different.

6.2.2. Qualitative responses

The received responses from the qualitative questions asked were few, however, they are still viable for minor analysis. They are translated from Swedish to English.

**Question 3.** What type of insight is gained by the information you are thinking of? To gain understanding of the information they have in mind, by query of what type of insights are gained from it (i.e. it has the intention of making it clear why the information belongs to a certain category).

"The customers experience of our technicians on location and the process” – Respondent 1

"How to satisfy the demand of the customer” – Respondent 4

The excerpts say the subscribers to AnalysSMS use their collected data to derive insights to improve customer service or customer relationships.

**Question 4.** “What are the limitations to how you currently derive insights from this information?” Limitations to their current way of analyzing the data.

“It is done manually, but it is not much work. Sadly we don’t receive much feedback.” – Respondent 1

"The summary could be simplified by being transferred to PowerPoint.” – Respondent 4

The first of the two excerpts in regards to the limitations of their data analysis of freeform feedback to have been conducted manually, though because there is so little data being collected it is not so much work.

The second of the excerpts would like the summary of the data to be transferred to PowerPoint; the reason being that AnalysSMS strictly presents the data in the web browser, thus makes it difficult to read the entries.
7. Results

The chapter presents the results of the analysis, investigating the business value of TM.

7.1. Information of Business Value

Having conducted text mining on the traffic school data and by using the concept matrix for business value of text mining, several problematic areas in the traffic school courses were revealed. Their issues have been highlighted through discovering patterns in the data, they have been validated with additional techniques such as correlation analysis, and been manually read by searching through the raw data, as a final validation, to be certain that the findings has given actual insights & knowledge about these aspects or features of the courses. These insights & knowledge, mediate the way toward the improvement of the courses for the traffic school and is therefore the business value that is derived from using text mining.

7.1.1. Information of Business Value derived

Time

The first insight that is derived is the one referring to the dissatisfaction toward the length of the courses. The insight could belong to the concept matrix category Customer Relationship, since needs, wants, and a possible Factor for Customer Satisfaction, are business value which the insight relates to; however, it is more relative to the category of Quality Control & Quality Assurance since the insight can be described as an insight of the subcategory Product Disapproval (cf. Concept Matrix Figure 1.), Service Disapproval might be more precise. How the traffic school would use this information in practice is a tricky subject. It is an assumption that the traffic school might be regulated by some domain-specific principles, in regard to how long the courses should be, and actions made upon this information might therefore be prohibited. If the latter is the case, then, the recommended course of action would be to revise the content of the course (e.g. be less repetitive, more stimulating), thus make the experience more enjoyable for the students’.

Room

The insight refer to the feedback given on the room aspect of the course. The analysis reveal the regulation of the heating is an issue and fixing it might improve the course experience. Similarly to the prior insight, this insight is hard to be classified into a specific category. It matches the business value of Product Disapproval, Detect Defects, Technical Issues, Factors for Customer Satisfaction, needs, wants. The information is of business value, as it can either belong to Customer Relationship or Quality Control & Quality Assurance, however, the author concludes it as being more relative to the Quality Control & Quality Assurance category.
Air & Ventilation

Though the patterns occurred in different keywords (air or ventilation), the terms are combined by the inferred relation of Air with Ventilation. The insight reveal the need to invest in a better air conditioning system. Additionally, the analysis revealed the reason to Room receiving complaints in regard to the regulation of the heating, being caused by the poor ventilation. It belongs to the category of Quality Control & Quality Assurance.

Hearing

The initial insight derived, gave the implication that the students’ difficulties in hearing were referring specifically to the teacher. However, it was revealed that the complaints in difficulties in hearing had many causes and were referring to varying aspects. For example, hearing difficulties could be caused by the teacher being too quiet, or that an interpreter speaking being too loud. Additionally, the insight also referred to the speakers (technology), which were used when watching some information or educational film, had too poor audio. Having the information that, it was difficult in hearing, did however, help in identifying these different issues. It belongs to the category of Quality Control & Quality Assurance, as it refers to technical issues.

Pain back neck chairs

The insights reveal the need to invest in more comfortable chairs to increase the course experience. The information says the chairs cause pain in the neck and back for the students, one of the excerpts comments on the courses being too long, and the chairs not being much of help by being so uncomfortable. It belongs to Quality Control & Quality Assurance.

7.1.2. Learning from Dataset

It possible conclude that the research objective 1 has been achieved; accomplishing objective 1 we’ve also learned and answered the related research question; furthermore, the fundamental problem to the study (proving the business value of TM) is solved.

Objective 1. “To conduct TM analyses, ensuring the business value of TM on a general type of Dataset.”

Research question. “What does TM analyses reveal about the investigated Dataset, and what does the analyses say about the business value of TM, given how well they agree with the concept?”

As presented in the above subsection, the business value of TM has been given cumulative evidence to its claim. Insights & Knowledge are derived from using TM, and with deductive/inductive reasoning, proven to belong to the categories of business value (as emerged in the literature), showing the concept matrix to be
working. The specific business value of each finding can be attributed to some elements found in the concept matrix: needs, wants service/product disapproval, technical issues, and defect detection. Additionally it was proven difficult to delineate each of the findings because of their interrelation among some of the categories, in this case Quality Control & Quality Assurance to Customer Relationship. The analysis of the findings categorized each into the category judged to be more appropriate. The reason why they are frequently placed in the category of Quality Control & Quality Assurance, is because the traffic school with certainty do not have recurrent customers, you take the courses, or a course, and do not need to return; therefore, it is arguably so that the information is more oriented toward bettering the experience for future customers, and thus the intention to increase the quality of the service hence them being placed into Quality Control & Quality Assurance. As the data is more oriented or framed toward extracting Quality Control & Quality Assurance related information for the courses, it affects the type of textual data that is collected. The survey could address this by purposefully querying the customers toward Customer Relationship-related questions; however, the necessity of such a query should be discussed.

7.2. SMS-survey

The SMS-survey had the purpose of accomplishing objective 2 followed by using findings of objective 1 and objective 2, to accomplish objective 3. Additionally the related research question of objective 2 can be answered, but the results are insufficient for drawing valid conclusions.

**Objective 2.** “To conduct an empirical investigation of companies, identifying general problematic business areas where business value from TM could be derived.”

**Objective 3.** “To compare the results from the business value as derived from TM the Dataset, with the identified business areas where such business value could be derived.”

**Research question.** “What are the general business areas where TM could be applied to derive business value from?”

The low responses for the SMS-survey left very little for analysis. What we can learn from the results is that Customer Relationship and Customer Service should be distinct categories, since the former span the entire interactions with the customer, while the latter strictly after purchase. Similarly, Quality Control and Quality Assurance are distinct, since they differ temporally; Quality Control taking part after a product/service has been developed, and Quality Assurance during development. Using the understanding gained by the TM case study, it is possible to conclude the described data to be viable for TM. However, considering one respondent having data where manual analysis is fine, applying TM to the respondent’s particular data would be unnecessary.
8. Discussion

8.1. Reflections on the research approach

To use text mining as a primary source for data collection, thereby having a basis by case study of business data, should be considered a compulsory when conducting an investigation into the business value of text mining; considering the infantile stages the research is currently in. To use nothing else other than secondary data to provide foundational evidence for the business value of text mining, would mean too much rationalization, and lack of empiricism weakening the claims. The latter, considering there is no prior business value in the literature (to the knowledge of the author), and thus only capable of making pure implications to business value; the argument is that the second-hand rationalized observations and conducting interviews upon these implications, would only prove guessed upon inauthentic hypotheses or concepts. This study has provided a starting ground by supplying a minor investigation of text mining in practical setting and creating a concept matrix; upon these findings further where corroboration is now needed before the relative hypotheses are created.

The SMS-survey did not discover any new categories, but they might be revealed in future research as they were probably impeded by the extent of the literature review. A more meticulous approach to the literature would therefore be needed. Furthermore, there is a need for a revision or to build upon the concept, since there are uncertainties to its definitions and the distinction of categories. Moreover, surely there are additional categories that should exists, for example, marketing or human resources and the ones already established could probably become more specific. The response frequency did not deliver in accordance with the expectations. 10 responses with some degree of nuance was expected. To increase the incentive for response, two movie tickets were raffled to the participants. Additionally the SMS-survey was described as demanding less than 5 minutes. Considering the extra efforts, the survey still failed. The reason why the response frequency was so low, is assumed to be the reason that the respondents in the companies are generally unwilling, uninterested, or find themselves preoccupied with more important tasks. However, the study was not dependent upon the result of the SMS-survey, as the empirical evidence gained from conducting the TM case study should be sufficient for an independent study. The latter considering the prior research content puts more emphasis on the activity of text mining textual data, and such a method for investigating the subject being more conventional. Strong empirical evidence from the SMS-survey would however enabled the TM case study empiricism to be used in comparison and draw a conclusion to the business value of TM more generally. In this manner, IP.1 who finds the technology interesting, could be more certain to what extent the technology is useful for their subscribers; additionally, companies who have access to similar data or would like to gain similar business value with the use of TM would find use of such insights.

In its entirety, the study could have been conducted with the use of a different method. One such method would have been to put the focus entirely on the case study
of text mining, with more emphasis on the company which the dataset belongs to. The method could use text mining to derive information of business value, and then evaluate impact of said insights & knowledge, as actions were made upon them. From then on the actual business value of text mining could be known as the initial state versus the changed final state of the courses are compared. It follows that such a method would be more dependent on time which was not possible for the current study. Additionally, the research would be affected by a common issue of case studies, that is, the issue of them being highly contextual, and having difficulties in generalizing the result. Considering the latter, the empiricism could be argued as stronger in regards to the business value of text mining, as it becomes more holistic by the data now being encompassed both temporally and spatially. The empiricism of the study that has been conducted can only claim to have been made on spatial data, since there were no temporal comparisons such as date of the written entries and their differences; or the difference (as discussed above) between the initial state and final state after actions have been made on the insights & knowledge. Instead the scope was put entirely on the traffic school, thus only spatial.

8.2. Discussion of results and recommendations in relation to the traffic school

By setting the focus of the study on how business value is derived with the use of text mining on an informational level, the analyses and results were very promising in regards to the empirical evidence of the case study. By utilizing text mining on textual business-data, the traffic school has learned what the different aspects of their courses are problematized and needs improvement. The utilization of text mining enabled access to previously inaccessible information. The latter accessibility, considering the size of the analyzed data. For a human to draw unbiased conclusions from such vast amounts of data is a too time consuming and close to impossible task. The derived insights & knowledge lead to different courses of action for the traffic school. Following these recommendations, the traffic school would thus be ensured a better course experience for the students and an improvement to the course quality.

- The courses should become shorter, as they are too long and lead to students being dissatisfied with the experience.
- The traffic school should be more attentive to how the students want the heating-regulation, since the heating regulation of the room is a negative aspect in regard to their experience.
- The students find the air conditioning to be bad. Investing in a better air conditioning system would better the experience.
- Students give different statements to hearing difficulties. The traffic school should invest in new speakers, so the students can hear what is said in the informational films. The teacher should take note of not being too quiet so the
students can hear more clearly. Since interpreters partake they should try to be quieter in order to not disrupt the learning experience of others, since leads to hearing difficulties for the other participants.

- The traffic schools has a need to invest in more comfortable chairs, since sitting in them for a longer time lead to pain and could affect the students ability to focus and learn. Overall it leads to a worsened experience.

The initial impression of the actions and improvements are them being too minor to derive any business value of substance. However, considering the value of word-of-mouth, if the customers are ensured that a product or service is to be trusted, a competitive advantage is sure to be gained in this regard. Compare this notion to how you yourself make decisions over your own purchases. We are often consumers of the products or services to which we have prior information (e.g. information that increase trust) which enables us to predict a product or service quality or the enjoyment of it and so forth.

8.3. Meta-Analysis of the TM-methods to derive information of business value

A meta-analysis was conducted where valuable findings in regards to the application of the TM-method and the execution of it are addressed. A multidimensional construct that could be referred to as “TM-methods to derive information of business value” is explained. It consists of three formative, underlying dimensions that are the factors with impact on the information that is derived. Here the intention is to provide additional transparency to the study, in regards to the TM-method and how future research should progress with the findings.

**Pattern recognition**

A necessity for the application of text mining-methods to derive information of business value, is its capacity/capability or sophistication to deriving patterns from any given data. The ideal is that the pattern recognition mediate as clear explanations as possible, i.e. them being self-explanatory. *Pattern recognition* is concluded to be a subcategory to the derived information being of business value.

Looking at the results from the text mining-method, and the analyses, show that the initially drawn conclusions are often correct, however, there were instances where the *effects* had various *causes*, meaning, it fails in presenting all of the relevant information that is available. More sophisticated techniques that could identify relationship between patterns, to find all *causes* would therefore be more self-explanatory, removing the need for additional validation, since the necessary information can be presented. An excerpt from the text mining analyses in relation to the pattern of “höra”:

“It is possible to identify what the causes are, however, there is too much variability in the data to be precise in knowing which cause is the most problematic”
Netzer, et al., (2012) identified co-occurrences in their data, as they revealed entities and related terms relationships to each other. It is evident that their text mining-method is more sophisticated, and it is possible to assume the same method could be used to find relative patterns in relation to a specific feature/entity, in the Dataset of the traffic school. With a similar method, one could for example select the feature of “höra”, and map its relationship to other terms by visualization, preferably other bigrams or trigrams. The text mining analyses was attempting a similar thing (Observe Figure 12.), as reported, the author experienced issues with encoding.

**Interpretability**

*Interpretability* refer specifically to the complexity in interpreting the presentation of the results that are derived from having identified patterns in the data, i.e. visualization or other techniques. The applied text mining-method used wordclouds for making initial inferences on the data, proving to be moderately accurate. *Interpretability* is affected by the *pattern recognition* since the latter lead to better possibilities for visualization. To improve *interpretability* the study employed visualizations such as, bigrams, trigrams and quadgrams. An excerpt from this study show explains the impact of N-gram visualization:

“*The pattern refers to pain in the back and neck, referring to the uncomfortable chairs. No correlation analysis should be needed, as the author find the N-gram quite self-explanatory*”

The quadgram (Figure 6), is more capable in capturing the semantics out of all employed techniques for *pattern recognition*. However, its weakness is a failure identifying the semantic relationships of different sentences.

The extracts presented in the literature review does not mention that much of *interpretability*, however, the study by Rose & Lennerholt (2017) give implications to the importance of *interpretability* with the following: “Interpreting text mining results is much more complex than indicated in the literature. Many of our results were hard to interpret, requiring leaps in sense making that even hardened qualitative researchers were unwilling to make.” The importance of *interpretability* can be understood by the following example:
The intention of figure 13, show the lack of interpretability as it attempts to cluster the bigrams into some sort of relationship based on their frequency. There might be some sort of insights to be gained, however, the interpretability is so low that the task would be a waste of time.

**Validity**

The issues of interpretability mentioned by Rose & Lennerholt (2017), lead up to the issues of validity, which could be explained by their statement: “In principle the application of text analysis is made insecure by the lack of an agreed, methodologically valid approach for the field.” The latter is relative to the experience of the author, in conducting the analyses. Considering the excerpt:

“However, only knowing what patterns exists and drawing shallow conclusions upon them, would not be enough to be certain on why the students give certain types of feedback or how improvements can be made. To validate the patterns and derive actual insights from them, should entail extracting some nuance of the data, which is possible with supplementary analysis using simpler qualitative techniques”

This study, entailed several steps required for actual validation of our findings. Similarly, Netzer, et al., (2012) from the literature review, similarly used traditional survey data in order to be sure that their findings had external validation.

There is a natural need for validation, however, the TM-methods would be of higher business value if the latter was addressed automatically by the TM technology. The additional steps required for validation, and to actually know what the findings mean, give implications to high distrust and insecurity towards the technology and its findings. The ideal would be the pattern recognition to have higher validity, where the initial conclusions are always or often trustworthy. If validation could be addressed automatically by the pattern recognition one would not need to search the raw data and expend time for making inferences in such a manner.
8.3.1. Toward a hypothesis for the business value of TM.

Through the conceptualization of the business value of TM (cf. section 2.2.) and the emergent results of the study, several concepts and their own inherent constituent components can be derived. Bhattacherje (2012, p.26) define constructs as: “abstract concepts specified at a high level of abstraction that are chosen specifically to explain the phenomenon of interest.” The author would describe it to consist of the central concept “business value” and attribute the concept to the concepts of: “TM-methods to derive information”, and “information”; to form “TM-methods for deriving information of business value” and “information of business value”. The concepts are multidimensional, meaning they consist of several underlying dimensions (Bhattacherje, 2012, p.44).

The author propose that the concept could enable an organization to measure or evaluate the TM-method to derive business information, if there was indicator to capturing the underlying dimensions of pattern recognition, interpretability, validity. They would first need to be operationalized, thus a need to develop indicators to capture measurement of each dimension (Bhattacherje, 2012, p.44).

Further measuring or evaluation would be possible with indicators to the business value of the information that is derived, for example operationalizing dimensions such as the impact of derived information, or how much information is being supplied. Furthermore, one could adopt the transactional, informational and strategical dimensions of IT-enabled business systems to measure the end-results. These measurement would make it probable to compare TM-systems through measurement or evaluation. The latter is all assumed by the author, as the measurable level or scale to variables such as “information of business value” or “TM-methods to derive information of business value” have no established metric.

Additionally, information which tells us whether a “room is green”, would not automatically entail business value is derived, since information is judged by it being new or useful to the receiver. Fruitless findings could imply that the method is unreliable for the current task or data source (minimal information is derived), and that there are methods more suited for better results. Similarly, the adopted methods for TM do not default to business value from information being derived, better TM-methods/systems would for example present information of higher business value; therefore, the TM method is a factor to the business value being derived and a correlation between the two exists, that is, a score of how much useful information is derived is measurable for a given TM-system.

Given the latter, it is thus possible to know that the two multidimensional dimensions, and inherent underlying dimensions, should be measurable on a scale, e.g. ordinal 1-5 or interval 1-100, to capture the value and some measurable difference to each. Nominal scale should exist for the dimension of TM methods for business value, where empirical backing exists as proof in it being capable to derive business value. As earlier stated, this should all be decided upon by the given context,
thus creating some limitations to the reach of the hypothesis in its capability to predict a particular physical reality. Creating a hypothesis could explain the causal relationship between the two dimensions since the chosen method for TM is a factor to the business value that is derived; this would better be addressed in a future study.

8.4. Contributions to Text Mining research

The Text Mining literature in its current body, is limited in terms of how well it explains how it achieves synergy in business settings. The study by Rose & Lennerholt (2017) is one of few works that provides practical explanations to how the technology can be utilized in this regard, though the intended contexts are different and leaves much for further investigation. The literature review collides with the text mining case study as the established concept matrix of the former, assists in deciding when information is of business value. The concept matrix is grounded on prior empirical evidence, and thus provides a way to investigate the business value of text mining on the informational level as well as help in grasping the concept. Furthermore, the concept matrix comprise the business domain (business domain where value is derived) and data source (the source from which the data originated), this level of detail is insightful as it can be used to know the exact type of data desired for certain analytical results. The latter indicates that data can become more insightful if they are designed appropriately toward the type of analysis one is expecting to conduct. If the intentions of the data collection is to know the technical issues of their product, a source should exist for such purposes. In accordance with the concept matrix, user-generated-content such as forums are shown appropriate to derive textual data for technical issues. In this regard the concept matrix is shown as having strategical capabilities, if data can be designed in this manner, the concept matrix supplies a meta-model for designing unstructured data for use in business settings.

The approach used, is argued as appropriate, considering the informational level is where the core value of the technology is derived. Furthermore, the author argue, this way of investigating business value of an IT to be a necessary improvement, as measuring the variables an emergent IT-enabled business systems will become increasingly difficult as they are certain to grow in size and complexity. The study therefore made its contributions to the research field, as it presents new insights that explain how text mining can be of advantage for businesses where the technology has synergy. Lastly, the study has supplied empirical evidence that the data source survey feedback, and the domain of traffic schools are contexts where business value with the use of text mining are derivable.

8.5. Scientific aspects

In the prior section, we discussed the dimension of validity and the general distrust toward the information that is derived from applying TM. It is therefore apparent that the research of the subject TM, should place an emphasis on the epistemological issues in using text mining. Addressing these issues are necessary as more analyzes are conducted on data such as Big Data. In the initial analyses it was possible to draw
semi-accurate conclusions based on the preconceived notions when looking at the visualized patterns, though, based upon the sophistication of the TM-method of this study, one could not claim to know or have actual knowledge since they were instinctual/intuitive guesses based upon an abstracted view from the actual source, i.e. the patterns said too little to be actually certain; additionally they were not capable of finding semantic relations (find all causes not a cause) to an issue, meaning we are more prone to error in our conclusions. However, as we can see with the example of Figure 6, given a more sophisticated NLP-technique (Natural Language Processing) the skepticism toward the results could be reduced (cf. nuance of quadgram contra bigram/trigram), and with a more precise pattern recognition that could find semantic relationships, perhaps all causes could have been included in this information.

8.6. Socio-ethical aspects

In considering the emergence of these new technologies that not only include text mining but would also comprise Data Mining, Machine Learning, Deep Learning, what springs to mind is how fallible they are in regard to the ethics of their use. It is not hard to imagine the dystopian potential of text mining, whose areas of application span a majority of the internet (everything you’ve ever written on the web or may write, be it that which is shared on social media or that which is supposed to be strictly private), there will come a time where for example something similar to a data warehouse or data lake might contain all of this sensitive information, and from there on, one could come to question the privacy of this information. Text mining is an enabling technology, the access to the data is the bigger issue. Would it be acceptable for any given entity to have access to this type of information? Ikeda, et al., (2013) could be used as an example. Based purely on the tweets of a person it is possible to derive demographics thus conduct profiling of individuals. Given access to more data and aggregating said data about any individual, the accuracy of the profiling can be increased. There are already businesses who might do the latter, though in the context of hiring (web crawling public textual data to find the best potential candidate). Section 8.4 hints at the collection of data should be designed in order to enable more insightful analyses, a relative socio-ethical aspect is the issue where the current manner for conducting business (by those who aim at becoming more competitive) are to build environments in which data from prospective consumers are constantly monitored where the end-intentions are to gain insights which increase our incentives for consumption to an extent where it’s closing in on brainwash, the latter since different personality types should be more receptive toward different types of directed marketing strategies. This would imply the technology invokes a use case that is highly pervasive in nature, where citizens are constantly being monitored, considering how much time of citizens daily lives are spent online. Additional concerns might be, as this work has already shown, how the transactional benefits of the TM technology made increased the efficiency of a manual process to such an extent that the need for human labor in the most time-
consuming area become redundant. In future settings, similar technologies might come to replace large portions of human workforce.

These notions are speculations made upon our current understanding and are not by principle strictly bad. Furthermore, there are still possibilities for the technologies to be used in contexts that are not detrimental to the society. Let’s consider how text mining could be used for the citizens and the betterment of society. An application for democratic purposes, where big data quantities of feedback is gained from the citizens. When feedback of the many can be analyzed in a qualitative fashion, we’d no longer be impeded by the current limitations of quantitative approaches to derive public opinions (excluding the trolls and sarcastic feedback). Though the text mining is still far from being capable of such a thing, this might be one of its future applications.
9. Conclusion and future research

With the emergence of technologies such as text mining, there is a growing importance to understand how the textual data of businesses can be harvested to derive information of business value. The capability of businesses to analyze their large amounts of textual data, not only externally (customers, social media data) but also internally (employee-data), opens up new paths to derive new insight & knowledge. There is little research conducted placed in a business setting that focus on how text mining can be of use for businesses, and most importantly what its part as a constituent component to business operations is or could be. This case study set out to prove the business value of text mining, and made its contribution by using the technology to perform analysis of survey-feedback-data from customers who’ve partaken in courses given by a traffic school. From this domain, information of business value was derived, specifically the distinct categories of: Quality Control and Customer Relationship. The questions asked were:

- “What does TM analyses reveal about the investigated Dataset, and what does the analyses say about the business value of TM, given how well they agree with the concept?”
- “What are the general business areas where TM could be applied to derive business value from?

To grasp the ambiguity of the concept “Business Value of Text Mining” a concept matrix was created. The result show it was from then on possible to categorize information into a specific business areas to decide when information was of business value. The result from the text mining revealed several patterns, and by further analysis mediated the way toward the improvement of the traffic school courses as insights & knowledge were derived from them such as: need & wants, customer satisfaction, service/product disapproval, detect defects, technical issues and more.

An SMS-survey was conducted though the results were not as desirable. The response frequency was too low, and limited the ability to draw any conclusion of substance. However, by its design the conclusion is made that the categories to the concept matrix should become more explanatory by making distinction between terms such as Quality Control & Quality Assurance, and Customer Relationship & Customer Service.

The results of the text mining analyses did provide patterns that were suitable for interpretation; however, there are additional techniques that could have been adopted if the complexity in their adoption was not impeded by the competency of the one conducting the analyses. R-Project does provide an archive with several open source packages which enables analyses of textual data for good end-results. Some of the techniques were fairly simple to use, as long as the pre-processing stages were
Pre-processing was the phase that demanded most time and was an important factor for decent results. With the black-box nature of the R-Project packages, the pre-processing steps has to make considerations to how the TM packages handles or view the data structure of the Document-Term Matrix. For the analyses to be correct (initially true) one has to make sure that it has been pre-processed for the right purpose. Additionally, encoding of the textual data was one of the limitations of the textual data. Using the method for Sentiment Mining, the tidytext package could not read the encoding of the characters in some of the data. Therefore, no relationships between sentiments and features could be extracted, which could have provided more insights & knowledge for the results. Issues of encoding should be addressed by the software that collects the data from the customers’.

As the quantity of available data increases, the data processing capabilities of a business should match its growth. Using text mining to make sense of large amounts of textual data will become a necessity for businesses, if their aim is to stay competitive. The business value of text mining is not only to know how a product or service is received by the customers’ (externally), but also help the businesses to understand themselves (internally).

As discussed in the sections 8.3 followed by 8.3.1 the text mining case study give several implications for research areas that could be explored to gain further understanding of text mining in a business setting. The result of this study experienced pattern recognition, interpretability and validity as the key factors to the information that is derived with text mining. If it is possible to know what factors have impact on the derived information of business value, then it is possible to evaluate or measure it in some way, to optimize a system. With the intention of implementing an automated text mining system, where these dimensions are not considered, the end business value would be less than it could be. This would need to be further studied as a hypothesis to explain how these factors affect the business value-phenomenon is further established.

Additional case studies in a business context should be conducted in future research since the existing literature focus on algorithms or techniques. This could help make improvements and expand the concept matrix created in this study. As mentioned in section 8.1, additional categories are sure to exist, this type of improvement would make the concept less ambiguous.
10. References


