CHAPTER 2
LITERATURE REVIEW AND THEORETICAL BACKGROUND

2.1. Literature Review

Social media can be a place to get feedback from customers about the services that a company has provided to its customers. However, companies have not always been able to process data from this feedback properly, and there is still confusion about the essential things to be used as a reference in business improvement. Burton and Khammash (2010) and Culnan *et al*. (2010) stated that the main advantage of using social media as a source of customer feedback comes from the fact that customer commentaries are posted on the Internet by the users for other users, creating an ideal environment for unobtrusive and non-intrusive research into customer mindset. Using data through social media will provide opportunities for companies to further understand what consumers need and expect of their products or services. Woodcock *et al*. (2011) stated that the knowledge gained on customer behavior, attitudes, and mood would help drive benefits throughout the value chain impacting on suppliers like forecasting demand and intermediaries like shaping in-store promotions. The high use of Twitter in the world with users sharing information and receiving real-time updates makes Twitter a suitable data source for companies. According to MacLeod (2010), its popularity appealed to many companies, seeking to not only reach masses of customers for sharing news and promotional materials but to, also, interact with them in real-time and address issues emerging pre-, during-, and post-service consumption.

In processing data obtained from social media, sentiment analysis is used so that the unstructured data can be processed by the company. Hence, they become valuable data for business improvement. Sentiment analysis is an approach of concept extraction resulting in opinion and sentiment from a natural language using natural language processing (NLP), text analytics, and computational methods (Miner *et al*., 2012; Liu, 2012). According to Ku *et al*. (2009) and Arnold (2011), sentiment analysis attracted significant attention in the last decade, promising to provide efficiency in analyzing subjective and unstructured online content, especially in the social media setting (i.e., products/service blog reviews, video posts, tweet feeds, etc.) Thelwall *et al*. (2010) state that the rationale for using Twitter feeds in sentiment analysis is that when a user responds to an event or emotion by tweeting, she/he demonstrates "information behavior," which essentially contains an affective component (e.g., judgments or intentions). He, therefore, reveals important information about users' opinions and sentiments on the tweeted topic.

In processing these data, a method such as Six Sigma or Lean Six Sigma is needed. According to Linderman *et al.* (2006), Six Sigma is a strategic initiative to drive profitability, increase market share and improve customer satisfaction through statistical tools. Fitriaty (2019) states that Six Sigma is a high process that helps the business focus on developing and delivering near-perfect products and services. According to Pavlović and Božanić (2010), LEAN was found by Taiichi Ohno in the 1950s. It started from the Toyota Production System with key aspects, including the never-ending quest for perfection, continuous search to eliminate waste, and the recognition and importance of employee contributions. Tenera and Pinto (2014) state that Lean focuses mainly on waste elimination, using visual and straightforward techniques whenever possible. Six Sigma and Lean methods can be used to get maximum results. Pavlović and Božanić (2010) state that many companies in different industries, both large and small, adopt Six Sigma and Lean as a common method to improve the efficiency of design, manufacturing, business processes, and intellectual property while reducing costs. These methods are integrated using the DMAIC (define, measure, analyze, improve, control) cycle to find a solution based on the data obtained. George *et al*. (2004) stated that DMAIC has proven to be one of the most effective problem-solving methods until now because it forces the teams to use the data to do help the solving of existing problems, see future opportunities and manage projects. In the DMAIC process, there is a process of defining Critical to Quality (CTQ). According to Tenera and Pinto (2014), One of the most critical steps of the LSS project aims to detect the Critical-to-Quality process factors (CTQs), considering the Customers' opinion. Reidenbach *et al*. (2002) said that CTQ is a concept commonly used in improvement projects, which serves to describe the different output characteristics of a particular process. CTQ factors are identified through stakeholders and market analysis in this research study, focusing on the VOC and taking real customer needs into account. VOC drives decisions and determines which people, products, or processes should be targeted for improved value courier (Found and Harrison, 2012; Narula and Grover, 2017; Martínez-Martínez *et al.*, 2018).

Companies can find VOC via Twitter and collect it. Tenera and Pinto (2014) said that the process's main critical quality factors (CTQs) were defined from VOC results. The VOC factors were then deployed into requirements, which were then translated into CTQ specifications that can be measured. Using CTQ helps us analyze qualitative data since CTQ uses direct data from customer comments in the form of an unstructured textual dataset. Moreover, CTQ helps preserve the sentiment data to do the sentiment analysis and feedback mining by considering the level of negativity or positivity of feedback (Aguwa *et al*., 2017). However, according to Antony (2006), the CTQ's are not always the same; they change according to the market dynamics, especially in service sectors where the customer expectations vary hugely. In addition, Dronamraju (2018) also states that the CTQ's should be critically analyzed and updated from time to time.

2.2. Theoretical Background

Theoretical background is provided to support the formulation of research planning and its implementation. The selected theoretical backgrounds are related to data and text mining, and business process improvement branches which become the focus of this research.

2.2.1. Big Data

Big Data is a term used for large heterogeneous sets of digital data acquired from the evolution and use of technologies that have been growing exponentially with time (Taylor-Sakyi, 2016; Riahi and Riahi, 2018). Big Data also can be distinguished by its characteristics. The early conceptualization of Big Data characteristic by Laney (2001) characterized Big Data using the 3Vs model: volume, velocity, and variety. Furthermore, this model was developed by Lomotey and Deters (2014) into a 5V model: volume, veracity, velocity, value, and variety. The figure of 5V model and its general explanation of each point can be seen in Figure 2.1.

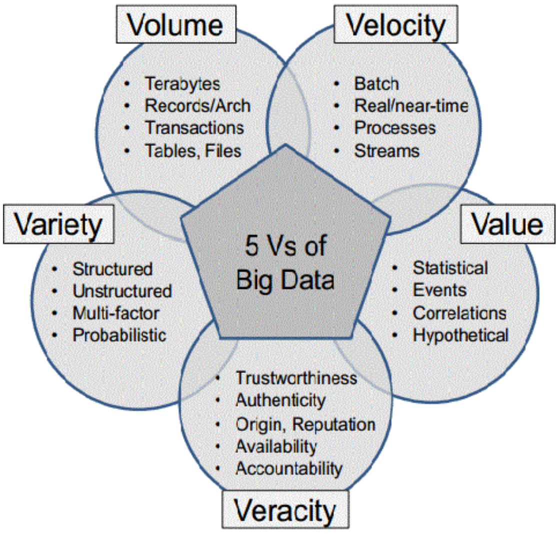

Figure 2.1. The 5V model that currently defines Big Data (Demchenko *et al.*, 2014)

Since the complexity exists in data, traditional techniques and/or algorithms can no longer handle the operation (Taylor-Sakyi, 2016). Big data aim to unveil hidden/ new patterns and relationship, and also to develop from a model-driven science model into a data-driven science model that can help organizations acquire specific knowledge from refined data (information) to make business decisions (Taylor-Sakyi, 2016; Chahal and Gulia, 2016; Rouse *et al.,* n.d.; Ackoff, 1989). The ability to achieve this aim is with big data analytics, which is an advanced method or technique of analyzing huge complex volumes of data (Elgendry and Elragal, 2014; Chahal and Gulia, 2016).

Considering the existence of data variety, as explained in Figure 2.1., the advanced method or technique of analyzing each type is divided into two: data mining and text mining. Text Mining patterns are extracted from natural language texts, *e.g.,* social media, while Data Mining patterns extracted from structured databases of facts, *e.g.*, spreadsheet and ERP system (Hassani *et al*., 2020; "What's the difference between data mining and text mining?", 2019).

2.2.2. Twitter Social Media

Social media stated by Merriam-Webster, 2004 is “a form of electronic communication (such as websites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages, and other content.”

With the definition and functions of social media, it can be used as a medium that allows the ease of findings public sentiment since social media can be a place for people to express their opinions. The availability of a lot of information on social media and the more common use of social media as a platform to connect business with the customers can make it easier to extract information related to public opinion on the product quality from certain company brand.

Twitter is one of the popular social media used to create status texts to provide information that will be placed on the Twitter timeline. In addition, tweets can also make users create a status about what the Twitter user is doing or feeling. Comparing social media existing on the Internet, Twitter is, currently, the platform that can provide the highest user opinions of various aspects, which can be useful for analysis purposes.

Based on Twitter Help Center Website (n.d.), some features and its definition existing on Twitter social media platform are:

a. Home

Users can see various posts / tweets from people who are followed.

b. Profile

Profiles serve to display various information shared by the account owner (e.g., tweets, media, tweets & replies, and likes) so that other users can see the activity of the account owner.

c. Followers

Followers are people who have followed certain users, so that the users who follow an account can see every activity carried out by the user on Twitter in the form of post sharing.

d. Following

Following is a state of users who follow other account owners who are is on Twitter, so that the users can see the activities carried out by the users who have followed on Twitter in the form of post sharing.

e. Tweets

Tweets is the slang for form of message posting in Twitter social media. The content that can be shared by the user is in the form of text, photo, video, link, geolocation, and others.

UNIVERSITAS ATMA JAYA YOGYAKARTA

f. Likes

Like button exist in the Twitter platform for the user to be able to like or show an appreciation of a tweet made by another user.

g. Reply

Reply button used to respond other people's tweet. This function is commonly used by the users to converse back and forth.

h. Retweet

Retweet button is used to spread or forward the tweet from another user to the users that follow the retweeter account.

i. Quote Retweet

Similar as retweet button, quote retweet is used to spread the tweet from another user to users that follow the retweeter account. The difference of quote retweet from retweet is that quote retweet has the ability to add one's own comment and/or media before retweeting.

j. Direct Message (DM)

Direct messaging allows twitter account owners to communicate privately to one or more targeted Twitter users.

k. Hashtag (#)

The hashtag is a symbol that resembles a fence which functions to make it easier for someone to search for a topic.

l. Mentions (@)

Mentions serve as the means by which Twitter users wish to have a conversation with the intended user account openly through the tweets that are made. Another use of mentions is to mark other user accounts that are related to the tweet content that the user has posted.

m. Trending topic

Trending topics have a function to summarize all the tweets that are currently being discussed by other users. What if there are many tweets that use hashtags (#) or certain words in a tweet. The Twitter topic trending feature can be adjusted according to geographic coverage, namely based on certain countries or worldwide.

2.2.3. Knowledge Discovery in Textual Database (KDT)

The process to acquire knowledge from Data Mining is known as Knowledge Discovery in Database (KDD) which is a broad process of finding knowledge in data, and emphasizing the "high-level" application of particular data mining methods (Hamilton, 2018). With the growth of the board exploratory in the type of unstructured data, the KDD process is, then modified into Knowledge Discovery in Textual Database (KDT). This modification is necessary to support the acquisition of knowledge from textual database using text mining (Jindal and Shweta, 2018).

- Text preprocessing
  - Syntactic/Semantic text analysis
- Features Generation
  - Bag of words
- Features Selection
  - Simple counting
  - Statistics
- Text/Data Mining
  - Classification-Supervised learning
  - Clustering-Unsupervised learning
- Analyzing results

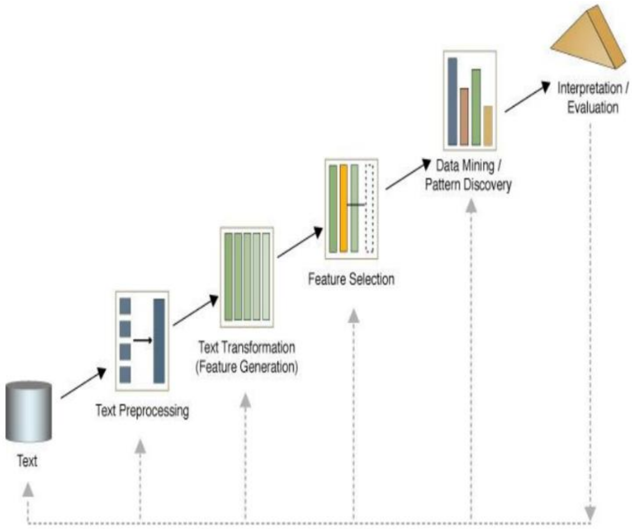
Interpretation / Evaluation
Data Mining / Pattern Discovery
Feature Selection
Text Transformation (Feature Generation)
Text Preprocessing
Text

Figure 2.2. Text Mining/ KDT Process (Liang, 2003)

The general process structure of KDD and KDT are still identical. However, the significant difference between KDD and KDT is the methodology used in executing each of the processes. Based on Liang (2003), several processes in acquiring the knowledge by KDT are text pre-processing, feature generation, feature selection, text/data mining, and analyzing results in which the details can be seen in Figure 2.2.

2.2.4. Sentiment Analysis

Sentiment analysis, also known as opinion mining, is one of text mining technique that aims to extract opinions and sentiments from natural language text using computational methods (Liu, 2015).

There are three fundamental grouping levels utilized: document-level, sentence-level, and aspect-level sentiment analysis. Document-level intends to characterize a supposition record as communicating a positive or negative conclusion or assessment. It considers the entire archive a fundamental data unit (discussing one theme). Sentence-level plans to arrange notion communicated in each sentence. Aspect-level intends to characterize the assumption as for the particular parts of the elements. The conclusion holders can contribute various thoughts for various parts of a similar element (Medhat *et al.*, 2014).

In the classification technique of sentiment analysis, three types of algorithm approaches used are lexicon-based, machine learning-based, and hybrid. The details of algorithms inside lexicon-based, machine learning-based can be seen in Figure 2.2.

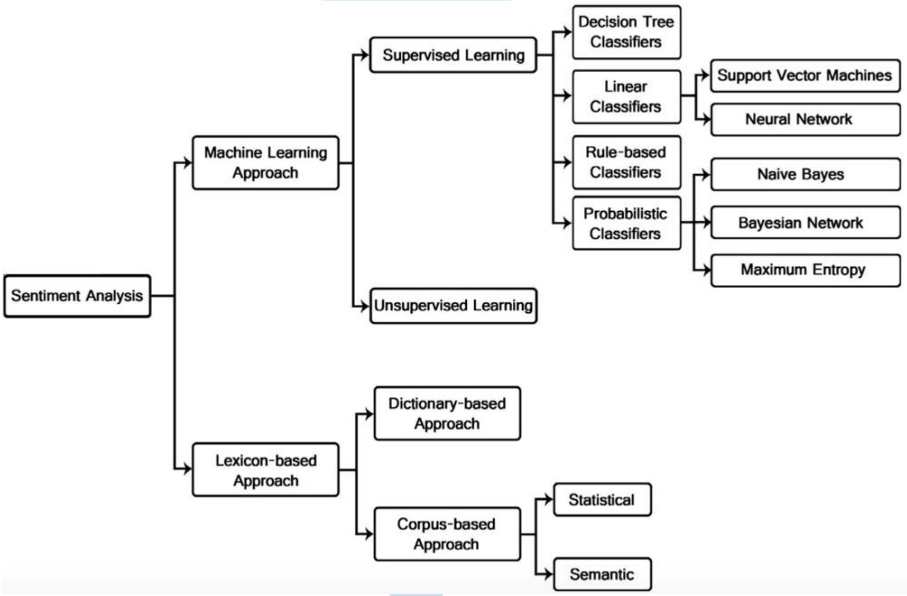

Figure 2.3. Sentiment Analysis Classification Technique

The lexicon-based approach depends on an opinion Vocabulary, a collection of known and precompiled notion terms. It is divided into word reference-based methodology and corpus-based approach, which utilize measurable or semantic techniques to discover sentiment extremity (Medhat *et al*., 2014).

The advantage of this approach is that it does not require any training data set. However, this approach has limitations in the expression detection (e.g., irony and sarcasm) and even though the approach can result in an outcome with high classification speed but the outcome still resulting low recall (Symeonidis, 2018; Isabelle *et al.*, 2019).

2.2.5. Comparative Study of Data & Text Mining Tools

Stated by Rangra and Bansal (2014) there are six best data mining open-source tools. However, in this research the considered tools to be used to aid this research are R and Orange. The selection of data science tools options for this research is taking into consideration of the three following aspects: free and open-source, ease of use and familiarity, and ability to achieve the desired objective of the problem-solving method. The tool that satisfies the most criteria is, then, selected as the tool used for this research. The comparison of R and Orange from the study conducted by Rangra and Bansal (2014) can be seen in Table 2.1.

Table 2.1. Comparison of Data Mining Method (Rangra and Bansal; 2014)

Nr.	Tool	Type	Advantage	Disadvantage
1.	Orange	Machine Learning, Data Mining, Data Visualization	a. Shortest script for doing training, cross validation, algorithms comparison and prediction b. Easiest data mining tool for learning c. Better debugger d. Simpler scripting data mining categorization problems e. Does not give optimum performance for association rules	a. Not super polished b. Big installation size c. Limited list of machine learning algorithms d. Machine learning is not handled uniformly between the different libraries e. Weak in classical statistics f. Reporting capabilities limitation in exporting visual representations of data models

Table 2.1. Comparison of Data Mining Method Cont. (Rangra and Bansal; 2014)

Nr.	Tool	Type	Advantage	Disadvantage
2.	R	Statistical Computing	a. Very extensive statistical library b. Ability to make a working machine learning program in just 40 lines of syntax c. Better integration of numerical programming d. Easier data import and export from spreadsheet e. Higher abstraction level f. Availability of numerous free packages, which provide all sorts of data mining, machine learning and statistical techniques. g. Support the intricate and complicated analyzes without deep knowledge of computing systems h. Suitable for analysis, graphics and software development activities of data miners and related areas	a. Less specialized towards data mining b. Existence of steep learning curve if not familiar with array languages

UNIVERSITAS ATMA JAYA YOGYAKARTA

2.2.6. R Programming Language & RStudio

R language is one of the most used programming languages used for big data analytics since its language and environment support statistical computing and graphics (Piatetsky, 2017; RProject, n.d.). In order to carry out data analytics, RStudio is used since it is an integrated development environment (IDLE) to develop R programming scripts.

To support sentiment analysis approach of user-generated Twitter database, RStudio equipped with several packages to aid the process of KDT. General packages in RStudio for Twitter database sentiment analysis are (Sutrilastyo, n.d.; Paradistia, 2019; Riadi, 2019):

a. rtweet,
b. twitteR,
c. RCurl,
d. corpus,
e. tm,
f. tidyverse,
g. tidytext,
h. textclean,
i. wordcloud2,
j. igraph and ggraph.

2.2.7. Voice of Customer

Voice of the customer is a term used to describe the process concluded in marketing research technique of capturing customer’s feedback, which consists of experiences, expectations, preferences, and aversions related to both quality and quantity aspects of current provided products/services (Powton, 2019).

Beneficial aspects that can be acquired for the business by considering this method are the ability to understand of the customer’s requirements in detail and, by that, it can be used as key inputs for the setting of appropriate design specifications for the new/improved product or service or as a highly useful insight for product innovation (Griffin and Hauser, 1991).

2.2.8. Business Process Improvement

Business process improvement is a systematic framework developed in management discipline used for the organization improve their efficiency, accuracy, and adaptability in the implementation of their business process (Harrington, 1991; Kissflow, n.d.; Andika *et al.*, n.d.).

UNIVERSITAS ATMA JAYA YOGYAKARTA

In general, the selection of a process for improvement stated by Harrington (1991) is as follows:

a. external customer problems and/or complaints,
b. internal customer problems and/or complaints,
c. high-cost processes,
d. long cycle time processes,
e. a better-known way (benchmarking, etc.).
f. new technologies are available,
g. management direction.

Stated by Harrington (1991), with measurable business process, benefits that can be gained are:

a. satisfaction of customer desire in appropriate time,
b. reduction of process cycle time,
c. reduction of Space requirements,
d. reduction of the number of steps and approvals,
e. reduction of noncritical output,
f. reduction of cost of the process,
g. reduction of cost of management.

There are various tools that can be used to improve processes in the business. The common tools/techniques among others are divided into three categories which are Lean, Six Sigma, and hybrid (Lean Six Sigma). Based on Rastogi (2020), the goal of Lean is to improve process performance through waste elimination (improving speed) and cycle time reduction while Six Sigma is to improve process performance in relation to what is critical to the customer (improving accuracy). To achieve both Lean and Six Sigma goals, combining both methods into a hybrid methodology (commonly known as Lean Six Sigma) is now regularly used to achieve more improved performance levels of the implemented improvement.

The most common tools used for Lean are: 5S, Value Stream Mapping, and Kaizen (PDCA Cycle). For the Six Sigma, the most common used tools are: DMAIC, DMADV, Cause and Effect Analysis, SIPOC Analysis, and Process Maps/ Process Flowchart.

With the development of business process improvement technique, the combination of Lean and Six Sigma is made into Lean Six Sigma (LSS). This developed technique overcome the barrier of objective to be achieved. With using

Lean Six Sigma, customer value of the products or services can be delivered through both efficient operations and quality standard (SSGI, n.d.). Lean focus on waste elimination which supports Six Sigma quality because waste elimination eliminates an opportunity to make defect while Six Sigma quality supports Lean speed since less rework means faster cycle time (Rastogi, 2020). For Lean Six Sigma (LLS) the tool/technique used to execute the process improvement is DMAIC which stands from Define, Measure, Analyze, Improve, and Control. The details of each steps in lean six sigma DMAIC framework can be seen in Figure 2.3.

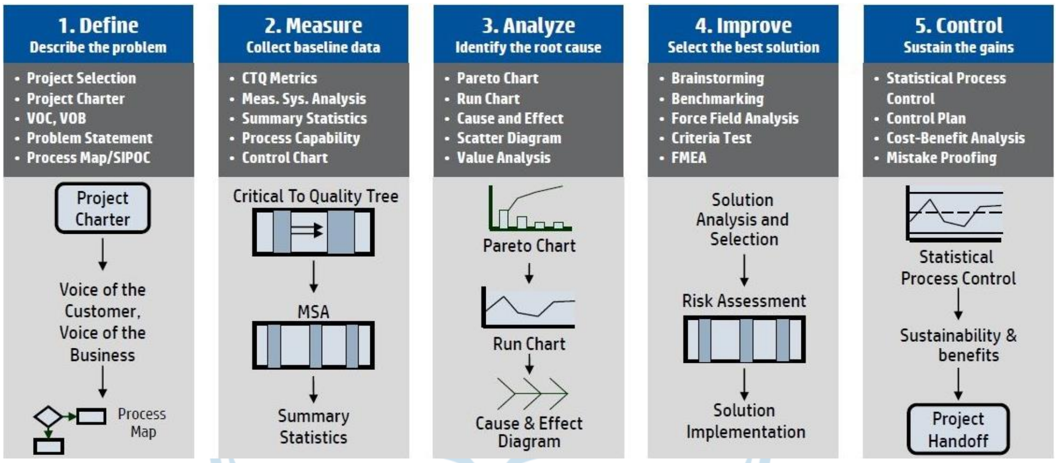

Figure 2.4. Lean Six Sigma Framework (Panat, R., 2014)