APPLICATION OF BIG DATA ANALYTICS FRAMEWORK FOR ENHANCING CUSTOMER EXPERIENCE ON E-COMMERCE SHOPPING PORTALS

by

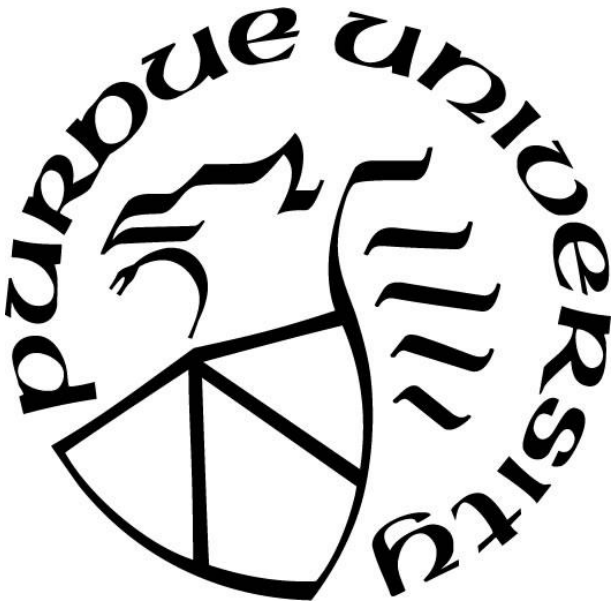
Nimita Shyamsunder Atal

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science



Department of Computer and Information Technology

West Lafayette, Indiana

May 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. John A Springer, Chair

Department of Computer and Information Technology

Dr. Chad Matthew Laux

Department of Computer and Information Technology

Dr. J. Eric Dietz

Department of Computer and Information Technology

Approved by:

Dr. Eric T Matson

Dedicated to my parents, Vandana and Shyam, for all their love; to my sister, Niki, and brother-in-law, Sameer dada, for being so supportive; and, to my beautiful baby nephew, Kabeir!

ACKNOWLEDGMENTS

I am extremely thankful to Dr. John Springer, my chair advisor and guide, for his guidance and support throughout this research. His advice and comments throughout the coursework have been extremely valuable.

I would also like to express my sincere gratitude to my committee, Dr. Chad Laux and Dr. J. Eric Dietz, for their constant cooperation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DMAIC – Define, measure, Analyze, Improve, Control

LSS – Lean Six Sigma

SS – Six Sigma

E-commerce (or e-commerce) – Electronic Commerce

LDA – Latent Dirichlet Allocation

GLOSSARY

Big Data → A large amount of data that can be accessed quickly; it is varied in nature; the value and meaning of the information contained within this data may differ over time, and it may not always contain the trustworthiness or ensure integrity (Blackburn et al., 2017).

Big Data analytics → According to Gandomi et al. (2015), big data analytics is a sub-process of the overall process of gaining meaningful insights from big data; it includes modeling the data, analyzing it, and interpreting the findings to extract insights.

Electronic commerce → Using telecommunication networks for performing business transactions, and for sharing and maintaining business-related information and relationships (Vladimir, 1996).

Six Sigma → According to Mehrjerdi (2011), it is a concept that can lead to quality improvement of the process or product by recognizing the connection "between the inputs to a product or process and the metrics that define the quality level of the product or process" (p. 80).

Sentiment analysis → According to Hipson (2019), sentiment analysis is a computational method that makes use of algorithms to detect and extract emotional content, i.e. positive and negative valence from the text.

ABSTRACT

E-commerce organizations, these days, need to keep striving for constant innovation. Customers have a massive impact on the performance of an organization, so industries need to have solid customer retention strategies. Various big data analytics methodologies are being used by organizations to improve overall online customer experience. While there are multiple techniques available, this research study utilized and tested a framework proposed by Laux et al. (2017), which combines Big Data and Six Sigma methodologies, to the e-commerce domain for identification of issues faced by the customer; this was done by analyzing online product reviews and ratings of customers to provide improvement strategies for enhancing customer experience.

Analysis performed on the data showed that approximately 90% of the customer reviews had positive polarity. Among the factors which were identified to have affected the opinions of the customers, the Rating field had the most impact on the sentiments of the users and it was found to be statistically significant. Upon further analysis of reviews with lower rating, the results attained showed that the major issues faced by customers were related to the product itself; most issues were more specifically about the size/fit of the product, followed by the product quality, material used, how the product looked on the online portal versus how it looked in reality, and its price concerning the quality.

CHAPTER 1. INTRODUCTION

This chapter provides a simple outline of the proposed research study. Different topics included within this introduction are the research questions, the statement of the problem, its significance, assumptions of this study, followed by its limitations and delimitations.

1.1 Research Question

This research study finds its motivation based on the following research questions:

- RQ1: Can a framework developed by using Six Sigma and Big Data Analytics principles be used for enhancing customer satisfaction in online shopping portals?
- RQ2: Which strategies could be implemented to improve upon factors affecting customer experience?

1.2 Statement of Problem

All organizations strive to improve customer experience on an on-going basis. Constantly growing competition and the ever-changing demands are the major catalysts that drive change and improvement in organizations. Continuous improvement in the processes is needed in order to enhance customer satisfaction. Six Sigma (SS) is a process improvement methodology, which aims to identify and then eliminate causes of the defects, thereby leading to improvement of the specific process. With the help of the application of a Six Sigma framework that integrates Big Data analytics methods and techniques, this study provides a basis for improving the process of analyzing customer reviews in an online shopping portal.

1.3 Significance

With the increase in usage of the internet all around the world, businesses have started developing and using online portals to increase their customer base. Lakshmi et al. (2016) have stated e-commerce to have a major impact on the global economy. E-commerce has multiple sub-domains contained within it, of which online shopping portals have gained a lot of momentum in terms of growth in the number of customers purchasing goods online. With the advent of these online portals, there is a huge amount of raw data getting generated which can be made use of for interpreting the online behavior of consumers (Lakshmi et al., 2016).

Fang et al. (2011) discussed how essential it is for organizations to pay close attention to their online review system and find correlations between product sales and online reviews. Resnick and Zeckhauser (2002) have stated that the transactions can be affected directly by the reviews provided by consumers online. There are several researches demonstrating the analysis of customer reviews using various methodologies using big data analytics, and so choosing the right framework can be extremely beneficial for organizations. Six Sigma methodology has been used in the industry for improving processes. Laux et al. (2017) developed a framework for Six Sigma DMAIC methodology and big data principles and techniques. This research study made use of this framework proposed by Laux et al. (2017), comprising of Six Sigma methodologies, for analyzing online customer reviews to understand factors that can impact customer experience.

Brandt and Reffett (1989), in their research, have stated that "ensuring quality of services that are regularly or continuously delivered to consumers" is indeed a challenge; their research also discusses that in order to improve quality of services, it is important to focus on problems that consumers face. This means that companies need to keep finding ways and means to improve service quality. Verhoef et al. (2009) looked at customer experience by taking a dynamic viewpoint and arguing that "prior customer experience will influence future customer experiences" (p. 31).

This is the reason why it is important to keep the social environment problems in viewpoint because past customer experience can affect future experiences (Verhoef et al., 2009).

Different researches have addressed the issue of improving customer satisfaction by application of different tools and techniques. With more demanding customers and even more competition in the market, retailers need to strive to come up with innovative ways that help with customer retainment, while also increasing their customer base (Srivastava and Kaul, 2014). Researcher did use existing methods and techniques, but along with this, they also proposed different industry models and frameworks to deal with this issue of better understanding online customer reviews in the e-commerce domain. To give an example, Bilgihan et al. (2016) developed a "theoretical model for a unified online customer experience" (p. 102). Min et al. (2018) have used Kano analysis, a Six Sigma methodology, in order to better understand and analyze customer requirements in online businesses while looking at the reviews provided by customers to ensure quality of service. Gupta et al. (2019) summarized the related literature on Six Sigma and big data analytics applications in order to understand how service industries and manufacturing industries can improve their processes.

While there are many techniques used for enhancing consumer's experience, application of a framework that combines the domains of Six Sigma and Big Data in the e-commerce domain helps improve a problem identified by performing in-depth analysis on data collected, in a structured format, and ensuring improvement in the quality of product/service – which many authors such as Brandt and Reffett (1989) have identified as a challenge. The study, when implemented by the online retailer, will help with the problem of identifying issues faced by the customer, as well as, identifying strategies which could be implemented to improve upon factors

affecting customer experience, thereby leading to customer satisfaction by continuous improvement – as stated in the Improve phase of this study.

1.4 Assumptions

- The third-party E-commerce data that will be used for the study is real commercial data
- The consumers have written their reviews honestly
- Increased customer satisfaction would reflect in more positive sentiments in reviews
- Increased customer satisfaction would reflect in higher rating
- Since the direct customers of the researcher of this study are the e-commerce organization, the researcher has assumed that the primary requirement of their customer is to enhance online customer experience

1.5 Limitations

- Final interpretations and findings may not be generalized for the whole population
- The study will only use the framework proposed by Laux et al. (2017)

1.6 Delimitations

- Proposal of a new framework applying big data analytics techniques and methodologies is not within the scope of this study
- Application of the recommendations that have been be provided after analysis in the improve phase is not in the scope of this study
- Since the improvement plan has not been implemented yet, the Control phase of the framework used does not specify how to go about implementing/standardizing improvements

CHAPTER 2. LITERATURE REVIEW

The contents of this chapter provide an overview of various topics affiliated with the research conducted. The literature related to the various topics affiliated with this research study has been presented in this chapter. Existing literature on Six Sigma projects, big data projects, and related methodologies has been discussed in this chapter. Furthermore, related researches that are done in the e-commerce domain, while also focusing on online shopping portals has been explored; customer experience in the e-commerce domain has been discussed with the help of related literature. Next, how Six Sigma can be integrated with big data methods and techniques has been discussed.

2.1 Big Data

Every organization from all the different industries produce massive amounts of information; the leaders and executives are faced with the same challenge of being able to use the information produced to its fullest value (Lavalle et al, 2011). Question is not just about the full usage of information, but it is also about gaining valuable insights to be able to compete in the market (Lavalle et al., 2011). The research conducted by Lavalle et al. (2011) was aimed at collecting information from "3000 executives, managers and analysts working across more than 30 industries and 100 countries" (p. 22) using a survey questionnaire, to help companies understand the importance of analytics. Based on the findings, Lavalle et al. (2011) noted that the organizations which have top performance tend to use analytics to a much greater extent in comparison to the other organizations.

Another important finding acquired by Lavalle et al. (2011) was that, based on their survey, organizations that strongly agreed on using analytics were likely to perform better in comparison

to the other organizations. Lavalle et al. (2011) categorized each of the organizations into three different levels primarily based on usage on the business information and analytics called aspirational, experienced, and transformed; where, aspirational organizations make limited use of analytics to justify the actions taken, experienced organizations do make little use of information and analytics to make decisions, and, transformed organizations are experienced and know how to make proper use of analytics "at organizing people, processes and tools" (p. 23) to gain competitive advantage in the market (Lavalle et al., 2011). Majority of research work by Lavalle et al. (2011) focusses primarily on how executives and leaders are wanting to adopt big data analytics to transform the way their organizations work, to gain better insights from the information generated, and to be able to better represent the insights, and so on. Lavalle et al. (2011) also compared the three different levels of capability organizations can be in based on the motive of organizations, their functional proficiency, business challenges, obstacles faced by the organizations, the way the organizations manage their data, the way analytics in used in the organizations, and so on.

Organizations are not just producing data of their own, but also capture information about their customers, their suppliers, and through daily communication and operations (Manyika et al., 2011). Big data can have immense amount of impact because it "can create significant value for the world economy" (p. 1) by not just increasing companies' productivity, but also affecting the customers by generating a surplus in economy for them (Manyika et al., 2011). The very reason for big data to be recognized by leaders from various industries is because "digital data is now everywhere" (p. 2); it is not just limited to a sector or economy (Manyika et al., 2011).

Manyika et al. (2011) characterized five ways to leverage the big data in a way that create value for the businesses and organizations; these broadly identified ways to manage, design, and organize their organizations include creating transparency, enable companies to experiment and

research more to improve their performance, understand requirements and taking actions based on the type of population targeted, enable better decision making capabilities with the help of automating the processes and using automated algorithms, creating new or enhancing existing services and products and processes, and so on. Manyika et al. (2011) studied five different domains and also discussed how making use of the data will not just help organizations to improve their products and services, but it will also enable companies to better understand their customers by finding patterns "in which customers, consumers, and citizens capture a large amount of the economic surplus" (p. 7). According to the research work done by Manyika et al. (2011), the authors discussed that amongst different sectors, there was about "60+% increase in net margin" (p. 8) of the financial value acquired in the US retail industry because of the usage of big data. Since the research work focusses on online retail, it is essential to understand the financial impact that big data has in this industry.

Over a while, there have been multiple techniques that have come into picture for analysis of big data "that draw on disciplines such as statistics and computer science" (p. 27); some of the techniques include A/B testing, classification, clustering, crowdsourcing, data mining, machine learning, natural language processing, neural networks, pattern recognition, regression, statistics, visualization, and so on (Manyika et al., 2011). Like there are various techniques, there are also multiple technologies used for analyzing, managing, manipulating data; these include Big Table, business intelligence (BI), cloud computing, data warehouse, data mart, MapReduce, Hadoop, R, SQL, and so on (Manyika et al., 2011). The research done by Manyika et al. (2011) focused on exploring in-depth the effect that big data on five primary domains including health care, public administration, retail, global manufacturing, and personal location data.

Of the five domains studied by Manyika et al. (2011), the authors also researched on the retail industry in the U.S. and how organizations within this industry are now adopting big data which has caused an increase in profitability. The authors discussed how this retail sector had the "potential to further increase sector-wide productivity by at least 0.5 percent a year through 2020" (p.64) because of digital data is being used and produced daily in an extensive amount between retailers and consumers (Manyika et al., 2011). The research pursued by Manyika et al. (2011) also showed how the U.S. retail sales - online and influenced via web – "forecasted to become more than half of all sales by 2013" (p. 66). Manyika et al. (2011) identified retail levers about big data and grouped the levers into different functions: Marketing function had the big data levers such as behavior analysis, sentiment analysis, cross selling, consumer experience, and so on; Merchandise function had the big data levers such as pricing and assortment optimization, etc.; Operations function had levers dealing with performance transparency; Supply chain function had big data levers such as inventory management, supplier negotiations, etc.; and, the new business models function included levers such as online markets and price comparison, and so on.

The research work by Shmueli (2016) focused on understanding the types of data, more specifically the behavioral big data which include data on human behavior, daily interactions, and so on. Shmueli (2016) described behavior type of big data as data that capture daily human behavior which included not just the actions taken by humans, but also the opinions and thoughts of humans about products, processes and services which have been self-reported. So, this type of data is primary dealing with human subject, since human behavior datasets that organizations use come from customers and employees gathered from different sources and using different tools (Shmueli, 2016).

There are number of researches done on behavioral big data, Shumeli (2016) concentrated on the research work done in the social sciences involving data about social issue and complex problems within. One of the domains taken into consideration by Shmueli (2016) was that of E-commerce organizations that made use of targeted marketing and behavioral data of their customers. Shmueli (2016) also gave an example of Target, which was a massive retailer company; they were in news because they made use of analytics to observe and recognize "changes in shopping habits" (p. 63).

A research by Borkar et al. (2012) summarized the history of systems used for management of big data, where both databases and newer systems were compared and contrasted "from an architectural perspective, looking at the components and layers" (p. 3) within the databases and the newer systems. According to Borkar et al. (2012), big data initially evolved from the need for organizations to maintain databases, data warehouses, data marts, and so on; after which big data emerged in the distributed systems and had an even larger impact. Different organizations came up with different systems of their own to deal with this big data, such as Google File System, Hadoop, Pig, and so on (Borkar et al., 2012). The authors, Borkar et al. (2012), compared the architectures of all these different systems developed by organizations to deal with big data; the authors also described the method applied to their "ASTERIX project at UC Irvine" (p. 12).

There have been fewer researches on "how processes can be institutionalized within organizations" (p. 2066) so that the impact that big data processes have can be improved (Saltz, 2015). The author, Saltz (2015), summarized works of previous researchers to understand different techniques applied for analysis purpose of the data at hand in a data science project; it included tasks performed such as information extraction, data modeling, using different frameworks and workflows like KDD and CRISP-DM, and so on. The motivation behind the research work pursued

by Saltz (2015) was essentially the fact that there was “no currently accepted process methodology for doing big data projects” (p. 2067).

The methodology applied by Saltz (2015) was to understand the requirement for having a process methodology, and how could organizations go about defining a method for finding solutions to specific questions. Furthermore, the author goes on to understand different domains, and how a project teams’ performance could be assessed based on usage of various models (Saltz, 2015). Some of the models explored included general model for measuring effectiveness of a project team, different information systems models to measure success, defining critical success factors, exploring maturity models (Saltz, 2015).

2.2 E-commerce

In general, information technology is affecting the lives of human beings daily, in a variety of phases and sectors; one such domain that is impacting customers is E-commerce (Gaffar Khan, 2016). The author, Gaffar Khan (2016), defined electronic commerce (E-commerce) as “the buying and selling of goods and services on the Internet” (p. 19), meaning, doing business electronically. Most organizations are molding the way they advance their business practices by also pursuing business online, and essentially using the internet to interact with their suppliers and consumers (Gaffar Khan, 2016). The conversations between customers and businesses have increased tremendously since the usage of the Internet and social media, in general, has increased; and developing nations are benefitting because of electronic commerce (Gaffar Khan, 2016).

Gaffar Khan (2016) discussed the benefits and challenges of E-commerce, along with researching on the E-commerce situation in Bangladesh specifically. The author, Gaffar Khan (2016), discussed some of the challenges that suppliers and consumers face while making use of E-commerce, which include: lack of cybersecurity whilst using websites; higher prices of internet;

lack of trust, especially when it comes to developing nations, with regards to online transactions or electronic settlements, and so on. Some of the benefits of E-commerce, for consumers specifically, as listed by the author include (Gaffar Khan, 2016):

1. Increase in comfort level and convenience since customers can purchase goods online at any time of the day without going physically to the store.
2. Reduction of delay in buying or selling of goods, which thereby causes time to be saved.
3. Comparison between prices and other factors of different companies can be done then and there itself.
4. Other customers' previous experiences about the brand or product can be found immediately in the form of self-reported review comments.
5. Helps improve the relationships between organizations and their customers
6. Helps increasing revenue of the organizations by reducing costs pertaining to maintenance and daily operations.
7. It helps increase organizations consumer base by raising customer retention; thereby developing the overall image and brand value of the organizations, and so on.

With the advent of big data and analytics, there has been a massive emphasis put on the application of big data and analytics in the E-commerce domain, however, in-depth exploration is still lacking (Akter and Wamba, 2016). The authors, Akter and Wamba (2016), have not only identified and characterized big data analytics features with respect to the E-commerce domain, but have also provided guidelines on how challenges pertaining to the application of big data in E-commerce can be handled, and further understanding the massive value that big data has on E-commerce. Since organizations, in order to survive in today's age of competition, always need to

be at the top of their game, adequate analysis of big data becomes essential in such E-commerce firms (Akter and Wambar, 2016).

The authors, Akter and Wamba (2016), summarized previous researches of big data analytics in E-commerce domain and segregated this previous research works on the basis of the primary themes of the study which include: identification of big data analytics in this domain; market segmentation; decision making, further leading to improvement in performance; new business models; infrastructure development, and so on. The authors, Akter and Wamba (2016), also discussed the different factors and nature of big data in analytics, which include volume, variety, velocity, veracity, and the description of each of these along with examples of studies. Organizations pursuing E-commerce primarily feel the urge of applying newer methodologies and techniques like big data analytics for "generation of economically worthy insights and/or benefits by analyzing big data" (p. 185) and thereby creating value (Akter and Wamba, 2016).

Of the many values that application of big data analytics has on E-commerce organizations, some of them include customer service and experience, supply chain visibility, security, detecting fraud, predictive analytics, and so on (Akter and Wamba, 2016). The authors also discussed the different E-commerce research streams, along with their relevant theories, and the future questions that researchers could take up and explore further (Akter and Wamba, 2016). Research questions within E-commerce are targeted towards different streams, customer experience on online shopping portals is considered for the purpose of this study. With the advent of the number of organizations opting for E-commerce, the number of customers using the internet for online purchasing has increased tremendously in recent years (Bilgihan et al., 2016).

By understanding and analyzing consumer behavior and their online shopping behavior, strategies could be proposed for enhancing the online customer experience (Bilgihan et al., 2016).

Organizations, nowadays, need to compete in a diverse and ever-competitive e-commerce environment; with this, the e-commerce organizations “have recognized the need to focus on providing a compelling shopping experience” (p. 103) since, according to the authors Bilgihan et al. (2016), not being able to provide the kind of online customer experience that is required would result in the loss of online revenue. According to the authors, Bilgihan et al. (2016), there have been very few researches conducted to investigate “online shopping experience that could assist to develop potentially important marketing strategies for companies” (p. 103). The researcher considers this as a proof statement showcasing the gap that is still existing in the related literature.

Bilgihan et al. (2016) have summarized previous research works stating the importance of having favorable customer experience, which can lead to an increase in customer loyalty. Bilgihan et al. (2016) focused on generating a theory-based model to enhance the aesthetics of the online environments in a way that it will create a positive customer experience in multi-channel interactions. But aesthetics are just one part, there are several other components that, according to Gentile et al. (2007), can be considered as dimensions of customer experience, these are: sensorial component – satisfaction with regards to aesthetics; emotional component – occurring through “generation of moods, feelings, emotions” (p. 398) and opinions towards a specific brand; cognitive component – engage consumer’s creative ability and the way they think; pragmatic component – usability of a product, also considers product lifecycle; lifestyle component – pertaining to behavior and beliefs of individuals; and, relational component – pertaining to the sense of community and togetherness with regards to a specific organization.

Hernández et al. (2010) stated that there are two groups of customers found online: the first group includes the individuals who consider purchasing goods for the first time online, whereas the second group includes customers who have at least purchased one product online. Customer’s

previous experiences can have a considerable effect on their future purchases, decision-making process, and e-customer behavior (Hernández et al., 2010). Rajgopal et al. (2000) stated that there can be various factors responsible for the success of e-commerce organizations including ease of use with respect to the website, range of products, quality of service, online communities, flexibility in terms of personalization of website, and so on. Based on the research findings of Rajgopal et al. (2000), online customer experience has been found to have a "viable long-term competitive advantage" (p. 24) on E-commerce.

Online shopping portals are a part of the E-commerce domain, wherein online shoppers tend to go through "people's reviews of a product to gauge their shopping decisions" (p. 42); so monitoring and analyzing product reviews are highly essential in order to understand opinions and sentiments of consumers (Zhang, 2008). According to the author, Zhang (2008), analyzing product reviews effectively can help with "polarity analysis and opinion extraction research" (p. 43). Zhang (2008) summarized the multiple pieces of research that have been previously conducted which have detected polarity and classified product reviews. Zhang (2008) also developed a "computational model that predicts a review's usefulness by exploiting its linguistic properties" (p. 44). Rose et al. (2011), in their research, developed a conceptual framework for understanding the factors that affect customer experience, and have stated that the subsequent consequences of attaining customer experience ultimately leads to customer satisfaction and re-purchase decision making.

Most e-commerce websites have recognized that consumers feel the need to go through product reviews of other consumers who previously purchased the product (Zhang and Tran, 2009). The authors, Zhang and Tran (2009), realized that it is difficult for the consumers to go through all of the reviews of a specific product, and they proposed "a model to discover helpfulness of online

product reviews" (p. 1). Agarap and Grafilon (2018) have also specified that knowing their customers' opinions pertaining to a specific product or service can help organizations come up with different marketing strategies, along with other things. For their research, Agarap and Grafilon (2018) made use of the dataset of Women's Clothing E-commerce Reviews by Brooks (2018). So, researchers have previously come up with different business models and used different methodologies and techniques for analysis of online consumer reviews.

2.3 Six Sigma

According to Dogan and Gurcan (2018), there are different types of quality improvement methods used for solving quality-related problems, these include "methods such as inspection, statistical process control, total quality control, zero defects, Kaizen and Lean Six Sigma (LSS)" (p. 943). LSS is a combination of Six Sigma and Lean methodologies, wherein, Six Sigma is used as a properly defined approach for problem-solving, and Lean includes reduction of non-value-added activities (Dogan and Gurcan, 2018). The researcher is focusing just on the Six Sigma part of LSS for their research, and not lean.

According to the definition of Six Sigma stated by Antony (2012), it is "a systematic, project-oriented, statistically based approach for removing defects from products, processes, and transactions" (p. 691). Six Sigma is a strategy applied in the form of methodology by organizations all around the world – not just in the manufacturing industry, but also in the service industry (Antony, 2012). Application of Six Sigma has proven to have improved the performance of the process at hand by not just improvement of the overall quality, but also reduction of costs, which has led to the increase in the brand value of the organization (Antony, 2012).

Antony (2012) performed SWOT analysis to understand "strengths, weaknesses, opportunities and threats" (p. 691) of Six Sigma by understand point-of-views of practitioners of

Six Sigma. According to Antony (2012), “very few empirical studies have been carried out on Six Sigma topics” (p. 697). Let us understand some of the common pointers with regards SWOT analysis performed, based on the findings of Antony (2012):

1. Strengths identified -
    - Emphasis is given on the quality of the product or process.
    - The impact on the performance of the process at hand can be shown by performing analysis.
    - Considered as a “data-driven methodology” (p. 695).
    - Emphasis on understanding business requirements.
    - Changes the way organizations function in terms of keeping a customer-focus mindset, giving importance to measurement, and thereby empowering employees, and so on.
2. Weaknesses identified -
    - It can be time-consuming, and it might need some amount of investment in terms of cost associated.
    - In terms of the process at hand, there may be a requirement of having “some level of organizational maturity” (p. 695).
    - Statistical knowledge can be considered as a must-have, and so on.
3. Opportunities identified -
    - The advent of information technology and the application of data analysis methodologies in all industries.
    - “Highly competitive market and demanding customer” (p. 692).
    - Considering manufacturing, service, or any other type of industry – Six Sigma “can be equally applicable in all processes” (p. 693), and so on.

4. Threats identified -
    - Misuse of Six Sigma is a potential threat that can affect the execution of the project at hand.
    - Because the process may require more time, organizations might not have enough patience or time at hand.
    - There is still a lack of training for Six Sigma certification, and so on.

2.4 Combination of Big Data and Six Sigma

According to previous researches, combining and then applying the two methodologies of Six Sigma and Big Data could lead to creative innovations, thereby causing an increase in economic productivity (Laux et al., 2017). The authors have compared these two methodologies in order to understand the differences or similarities existing in the principles of Six Sigma and Big Data. There have been previous studies conducted, wherein authors have come up with frameworks "to take the understandings from descriptions of SS and Big Data" (p. 668) in order to combine and explore further (Laux et al., 2017).

The authors, Laux et al. (2017), have also described two methods that can be used while developing an improvement approach, where the process at hand can be looked at in a top-down or bottom-up approach. According to Laux et al. (2017), the top-down approach consists of finding whether improvement in the process will "help us meet our business goals" (p. 668), whereas, the bottom-up approach consists of understanding if the process performance will improve if the problems within the process at hand are fixed. Six Sigma consists of different methodologies, of which one of the prescriptive frameworks is the DMAIC methodology which includes "define, measure, analyze, improve and control" (p. 666) phases that can be used for finding the root causes

(Laux et al., 2017). The authors, Laux et al. (2017), proposed a framework that combines Six Sigma DMAIC methodology with the techniques from big data as shown in Figure 2.1.

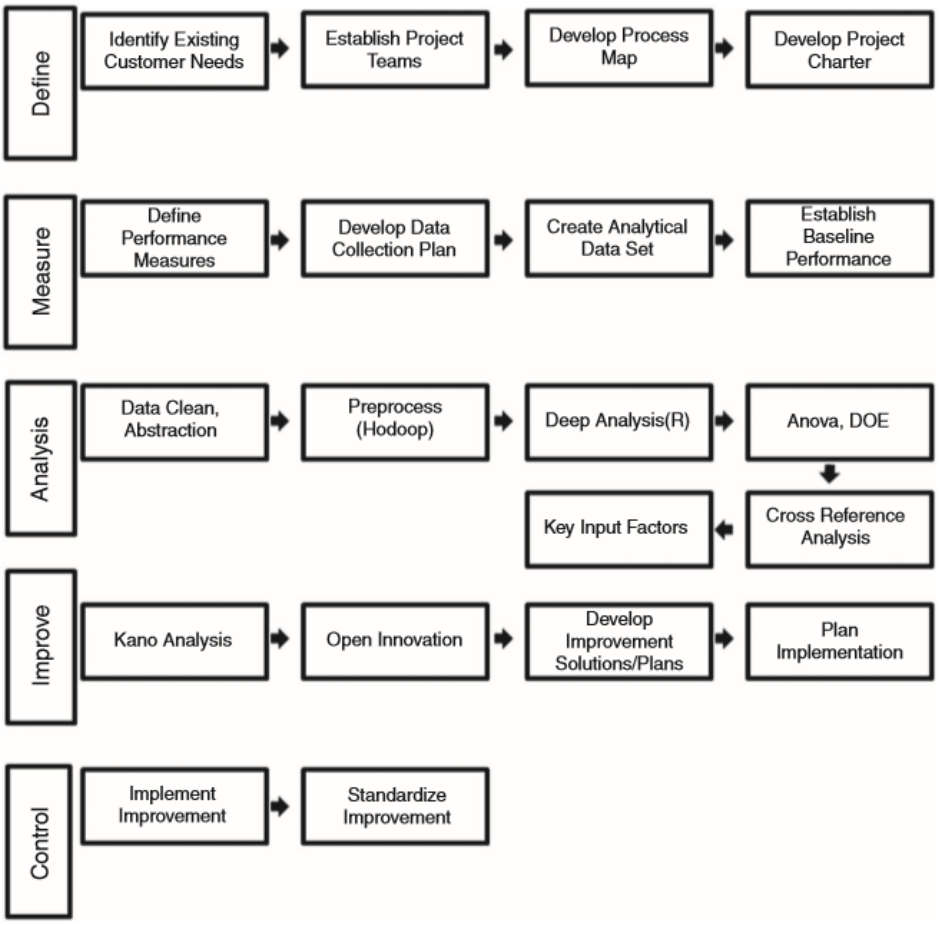


Figure 2.1 Framework by Laux et al. (2017) combining Six Sigma and Big Data

Laux et al. (2017) proposed a framework that used the DMAIC methodology of Six Sigma, and it involved techniques of Big Data. According to Laux et al. (2017), the framework in Figure 2.1 includes Big Data techniques put in each of the phases of the DMAIC methodologies, wherein:

1. Define Phase consists of – Identification of requirements from the customer's standpoint. Data collection should be done from appropriate sources in a way that will tell us the voice of consumers. Collecting customer data is essential can help make the process at hand more

"customer demand-driven" (p. 670). Based on the framework, other steps in the define phase include developing a process map and a project charter, and so on.

2. Measure Phase consists of – Performance measures and baseline performances can be identified in the measure phase of a traditional DMAIC methodology; however, the framework by Laux et al. (2017), has additional steps added in the measure phase. According to the authors, Laux et al. (2017), "the company should take historical data set as the first choice" (p. 670), the next steps involve filtering data to create an analytical dataset.
3. Analyze Phase consists of – Making use of software tools like R in order to handle and process the big data at hand and visualize it further. Laux et al. (2017) stated that this step can also involve using "traditional statistical tools in SS, such as ANOVA or DOE to explore the possible causes and effect relationship" (p. 671).
4. Improve Phase consists of – Kano analysis can be done, which can assist in realizing the expectations of consumers which can keep changing over time. In this phase, Laux et al. (2017) stated that recommendations for improvement required could be stated, and accordingly a plan could be made in order to implement the recommendations suggested.
5. Control Phase consists of – According to Laux et al. (2017, both Six Sigma and Big Data methodologies "can mutually benefit and strengthen the power of constant improvement" (p. 672) by further ensuring the sustainability of the project at hand. This last phase includes implementing the recommendations stated in the previous phase.

2.5 Six Sigma in E-commerce

According to the definition of Six Sigma by Behara et al. (1995), it is a way of calculating and understanding the probability that organizations can produce a product or service having zero

defects. This methodology was originally applied just to manufacturing organizations, organizations have "extended the concept of zero defects, measured by Six Sigma, to customer satisfaction" (p. 1) over the years (Behara et al., 1995). The authors, Behara et al. (1995), presented a case study wherein a high-tech manufacturing organization wanted to use Six Sigma on its customer satisfaction measurement "to examine the impact of customer expectations on the company's strategies for improving satisfaction" (p. 1). The authors, Behera et al. (1995), have described the process of customer satisfaction to be a multi-stage process in terms of having multiple facets such as customer service, product or service delivery, product quality, and so on, thereby making it "more difficult to reach a level of Six Sigma in the customer satisfaction arena" (p. 3).

Laux et al. (2017) gave a pseudo example of application of their framework in the Higher Education Institution (HEI) domain which, according to the authors, can lead to a much more integrated improvement in the overall system. While Laux et al (2017) proposed the framework, the researcher found motivation for this research based on the gap in this area of literature since the framework had not been previously tested and applied to the E-commerce domain specifically. While Behera et al. (1995) has used Six Sigma to improve customer satisfaction, they have not used big data techniques to enhance the effect of Six Sigma. Based on the literature review that the researcher had done, they found a gap in the literature wherein, while researchers have used only Six Sigma techniques or only big data techniques for analysis purpose in the e-commerce domain, the researcher did not find an application of framework or business models which uses big data techniques in Six Sigma's DMAIC methodology to analyze online customer reviews in the E-commerce domain, more specifically pertaining to the online shopping portals; the

researchers took this point as the primary motivation, and they have addressed this gap through this research work.

CHAPTER 3. METHODOLOGY

3.1 Research Methodology

According to Akter and Wamba (2016), the rise in the usage of big data and analytics has put tremendous emphasis on its application in the E-commerce domain, however, in-depth exploration is still lacking. While e-commerce is a very large domain, the researcher focused just on the online shopping sector, and within that, the online customer reviews about clothing and apparel products. According to the authors, Bilgihan et al. (2016), there have been very few studies conducted to investigate "online shopping experience that could assist to develop potentially important marketing strategies for companies" (p. 103). Bilgihan et al. (2016) have summarized previous research works stating the importance of having favorable customer experience, which can lead to an increase in customer loyalty. Knowing and understanding their customers' opinions pertaining to a specific product or service can help organizations come up with different marketing strategies, along with other things (Agarap and Grafilon, 2018).

Laux et al. (2017), through their research work, worked towards recognizing whether Six Sigma methodologies could be reinterpreted in such a way that it "holds promise through a theory-building idea" (p. 663). The authors, Laux et al. (2017), proposed a framework that was based on the DMAIC methodology of Six Sigma, wherein, each of the different phases of the DMAIC approach consisted of Big Data techniques. One of the main questions that Laux et al. (2017) focused on was to understand if big data is "the next area for SS practitioners to adopt" (p. 663). The primary methodology of this research study follows Laux et al. (2017), wherein, the researcher took the framework proposed by the authors and applied it to the e-commerce domain; validation of the framework is not within the scope of this research.

The nature of the study at hand is exploratory since, according to Yu (2017), exploratory data analysis is a systematic approach by means of which the problem could be looked at from different perspectives. In general, exploratory data analysis can consist of various taxonomies including big data analytics techniques, "data re-expression, resistant procedures, and data visualization" (p. 2) amongst others (Yu, 2017). With the application of exploratory data analysis, the aim of the researcher was to explore the data at hand using different techniques till patterns can be recognized and a story can be formed based upon it (Yu, 2017).

Furthermore, the study has used both qualitative, as well as, quantitative methods; it is therefore a mixed methods research (O'Cathain et al., 2007). The contents of this chapter provide an overview of the methodology approach that has been taken for this research study. Since the whole idea and intention was to test the DMAIC framework proposed by Laux et al. (2017), different analysis and design procedures pursued within each of the phases of the DMAIC framework have been discussed in the section stating the procedure of the research study within this chapter.

3.2 Procedure

Laux et al. (2017) proposed a framework consisting of both, Big data and Six Sigma, methodologies. The authors stated that "more than 55 percent of" (p. 667) projects consisting of only big data techniques fail; some of the primary reasons for failure to occur include "lack of understanding how to use analytics to improve organizational processes" (p. 667), not having the right skillsets required or the not using appropriate framework or techniques, amongst others (Laux et al., 2017). By combining the Six Sigma DMAIC framework along with big data techniques, according to Laux et al. (2017), helps solve problems in a systematic approach since the "principles

of SS and Big Data may share synergy" (p. 667). The different phases of the DMAIC framework proposed by Laux et al. (2017) have been applied in the following way:

1. Define Phase –
    - Identification of the existing customer needs: Since the researcher is acting as a consultant for the E-commerce organization, the customer is the organization itself. For this research study, the customer is the E-commerce organization whose customer reviews dataset the researcher worked on. Since the customer is the E-commerce organization, an assumption made was that the customer's main requirement is to enhance their customer's (or consumers) experience.
    - According to Cudney (2012), the project charter is something that is developed by the development team at the beginning of the project, in order to understand and limit the scope of the project to a point that it is acceptable by all stakeholders involved. Based on contents of a project charter stated by Cudney (2012), for this research study, the researcher had the following deliverables as part of the project charter: a problem statement stating the business case at hand; list of stakeholders, along with the consultants, project leaders, Black Belts, and so on; a goal statement consisting of what is important to the customers; a motivation or opportunity statement that tells us what is critical to satisfaction; describe project deliverables in the form of a project plan, and scope statement.
2. Measure Phase –
    - For developing a data collection plan, Laux et al. (2017) suggests that historical dataset should be chosen because most of the times "the cause of the problem already exists in historical data" (p. 670); so the researcher did not have to spend extra resources to

collect real-time data since the researcher used a dataset published by Brooks (2018) which consists of online customer reviews on a women's clothing e-commerce website.

- Description of the dataset:
    - It consists of 23486 rows and 10 feature variables, wherein each of the rows within the dataset represents a review of one corresponding customer, and so on (Brooks, 2018).
    - The 10 featured variables within the dataset include (Brooks, 2018): Clothing ID – which is an integer variable given to a particular clothing item; Age – which consists of the age of the customer who wrote the review; Title - describes a brief title of the review; Review Text – which consists of textual review given by the customer, in the form of a string; Rating – a score that customer gives to a specific product, rating from 1 (worst) to 5 points (best); Recommended IND – which specifies if the customer had recommended the product or not, represented by 1 and 0 respectively; Positive Feedback Count – states how many consumers have given a positive review; Division Name – division name of the specific product; Department Name – department name of the specific product; Class Name – class name of the specific product.

3. Analyze Phase –
    - A major feature that was used is the review given by the customer in the form of text; text analysis was performed extensively in this research study. R software, as recommended by Laux et al. (2017), was used for analysis purposes.

- For preprocessing, the following process (Meyer et al., 2008) was referred: preparing the data, cleaning it, and then general preprocessing; followed by conversion of text to the corpus. To clean the data, the following process was referred to: converting all text to lower cases, and then removing all punctuation, followed by removing stopwords, and stemming the document to clean the data.
- The intention of this research study was to analyze the text which, along with other processes, consisted of performing exploratory data analysis by performing sentiment analysis on text data, using the concept of the linear regression model to check which factors have the most impact on the sentiment score, and also comparing the reviews of individuals who have recommended a product with the ones who have not. The intention was to detect pain points and understand areas of improvement.

4. Improve Phase –
    - Analyze phase results gave the pain points or the issues faced by the customers; based on these pain points identified, recommendations were made by the researcher in the areas of improvement.
    - Methods to plan this implementation were discussed in this phase.
5. Control Phase –
    - Implementation of the improvement plan was out of scope for this research, hence in this phase, the researcher has only recommended how the recommendations could be implemented, and how the organization could sustain and maintain the improvement going forward.

3.3 Sentiment Analysis

This research study made use of sentiment analysis in order to better understand the reviews by customers. The following steps were followed in order to perform sentiment analysis in R would be as stated below (Littlejohn, 2018):

- Tidytext package by Silge and Robinson (2016) was installed in R
- AFINN lexicon was used; specifically, because this would help with assigning sentiment scores to reviews
- Review field containing review texts were tokenized
- Based on the scores given to words in the afinn list, a mean afinn score of the tokens within a review was calculated to get a sentiment score for that specific review
- Further, the polarity of customers was observed on the basis of the sentiment scores; meaning, if the sentiment score (mean afinn score) of the review was above 0, it was positive; if the sentiment score of the review was below 0, it was negative; and, if the sentiment score of the review was equal to 0, it was neutral.

3.4 Linear Regression Modeling

The next step of the study was to generate a linear regression model. The following steps were followed in order to perform regression analysis:

- The researcher wanted to understand the impact of different variables on the sentiment score, to see what affects the opinions of the users the most, so the researcher used the linear regression modeling concept for this purpose.
- Simple linear regression model, as shown by Fox (2003), was generated by making use of the lm() function in R; wherein, the sentiment scores given to reviews acted as the

independent y variable, whereas, the Age, Recommend, Rating, Department acted as the x dependent variables.

3.5 Topic Modeling

Finally, topic modeling was performed; and the following steps were performed for generating:

- The output of the regression model generated showed that the Rating field, amongst all fields, had the most impact on sentiment score.
- So the researcher took the subset of data containing reviews which had been given a lower rating of a 1, 2, or 3, and used Latent Dirichlet allocation (LDA) to generate topics, with the intention of understanding the main issues faced by the customers who have given a low rating to the product (Calheiros, 2017).
- For this, the researcher used the topicmodels package by Grün and Hornik (2011) in R and generated two topics by referring to the code published by Soltoff (2020).

CHAPTER 4. RESULTS

The contents of this chapter provide results of the analysis done using different methodologies and techniques. The researcher made use of the framework proposed by Laux et al. (2017) to systematically use the techniques and methods for analyzing and going through all the phases of the DMAIC methodology. This chapter has been divided into each of the five phases of the DMAIC methodology framework proposed by Laux et al. (2017) which include the Define phase, Measure phase, Analyze phase, Improve phase, and Control phase, along with the approaches and results of each phase; the final section of this chapter also shows the brief description of the final framework with methods and techniques.

4.1 Define Phase

The first phase of the DMAIC methodology, according to Laux (2017), is the Define phase; this phase consists of identification of customer requirements. In this study, the researcher considered the e-commerce organization, whose dataset is being utilized, as the customer/primary client. For the purpose of this research, the researcher made assumptions for the e-commerce organization's requirements. The researcher described the requirements of the Define phase in the form of a project charter.

4.1.1 Project Charter

In their research, Cudney (2012) stated how a project charter could be developed for different projects; also, what all factors could be included in a project charter to make it more customized for that specific project. Hayes (2000), on the other hand, also specified different sub-parts that should be included in a project charter, primarily on the basis of evaluating different

project charter templates already existing at that time including the template of the charter given by Tryon and Associates. Based on these two studies, the researcher examined and considered necessary sections to generate a project charter for this research, which is summarized in the Table 4.1 below.

Table 4.1 Project Charter

<table>
<tr><th colspan="2">PROJECT CHARTER</th></tr>
<tr><td><u>Problem Statement</u>
Application of framework proposed by Laux et al. (2017) to an e-commerce organization</td><td><u>Opportunity Statement</u>
Online shopping portals (Lakshmi et al., 2016)
Past customer experiences (Verhoef et al., 2009)
Recognizing KPI – customer satisfaction (Rose et al., 2011)</td></tr>
<tr><td><u>Goal Statement</u>
Collect online customer reviews of the e-commerce organization
Explore dataset with the intention of improving the KPI to identify major issues faced by customers
Suggest strategies to improve upon factors affecting the KPI</td><td><u>Scope Statement</u>
Core Process: Analysis of online customer reviews
Start of Process: Test framework, following DMAIC methodology, proposed by Laux et al. (2017) on the online customer reviews dataset
End of Process: Interpret the results obtained on each stage of the analysis to answer the research questions</td></tr>
<tr><td><u>Project Deliverables</u>
Project Charter
Data Collection Plan, Basic Performance Measures
Preprocessed data; Outcome of exploratory analysis, sentiment analysis result, comparison of effects of different factors on sentiment score, linear regression model, topic modeling
Improvement Solution/Plan
Suggestions on how organizations could sustain the improvement</td><td><u>List of Stakeholders</u>
Project Sponsor: E-commerce Organization
Project Champion: Dr. Springer
Black Belt: Nimita Atal
Consultants: Dr. Laux, Dr. Dietz
Target Audience: E-commerce organization and their customers</td></tr>
</table>

This includes the following sections:

Problem Statement –

To test a framework, proposed by Laux et al. (2017), consisting of Six Sigma and big data techniques on an e-commerce organization online customer reviews dataset, with the aim of analyzing how customer experience could be enhanced for the specific e-commerce organization. While there have been researches performed on how the online customer reviews pertaining to e-commerce organizations or retail stores could be analyzed, there is no existing research work focusing on the application of framework proposed by Laux et al. (2017) which combines Six Sigma and Big Data methodologies to an e-commerce organization; furthermore, this research aims to enhance the online customer experience on the basis of the framework and by using the customer reviews dataset at hand.

List of stakeholders –

Project Sponsor: E-commerce Organization

Project Champion: Dr. Springer

Black Belt: Nimita Atal

Consultants: Dr. Laux, Dr. Dietz

Target Audience: E-commerce organization and their customers

Goal Statement –

- Collect data consisting of online consumer reviews of an e-commerce organization
- Clean the dataset acquired for further analysis
- Conduct exploratory data analysis on the acquired dataset
- Generate sentiment score for each row of the dataset- wherein each row represents a single customer review

- Analyze the effect of other feature variables on the sentiment in order to understand which features have an impact
- Identify areas of improvement for the organization to improve the online experience for their customers
- Provide recommendations for improvement in the online customer experience process

Opportunity Statement –

Online shopping portals keep gaining popularity and attracting audiences from different brackets every single day. So, E-commerce organizations, or organizations managing these online shopping portals, need to keep finding and implementing newer ways of improving the quality of their service, as stated by many researchers (Brandt and Reffett, 1989). Organizations are realizing how important it is to look at past customer experiences since that can have an impact on the future experiences (Verhoef et al., 2009). One of the biggest domains that can help with this process is making use of big data; Six Sigma is another methodology that can be instrumental in analyzing the requirements of customers. The framework proposed by Laux et al. (2017), which combines both of these domains, can help with not just understanding requirements, but also analyzing behavioral aspects pertaining to the customers; moreover, with this framework, the organizations can set goals aligning with both the Big Data and Six Sigma domain, and have a structured methodology for analysis purpose in place.

Project Deliverables –

This would include the deliverables attained in every step of the DMAIC methodology, which would include the following deliverables:

- Define: Project Charter

- Measure: Data Collection Plan, Basic Performance Measures
- Analysis: Preprocessed data, Outcome of exploratory analysis, Sentiment analysis result, comparison of effects of different factors on sentiment score, Regression model, Topic Modeling
- Improve: Improvement Solution/Plan
- Control: Suggestions on how organizations could sustain the improvement

Scope Statement –

- Core Process: Analysis of online customer reviews
- Start of Process: Test the framework, following DMAIC methodology, proposed by Laux et al. (2017) on the online customer reviews dataset that is considered for this project
- End of Process: Elaborate the results obtained on each stage of the analysis to answer the research questions pertaining to this research work
- The scope of this project does not include proposing a new framework, nor does it include implementation of the improvement plan suggested in the Improve and Control phases of the DMAIC methodology used

The scope of this project and the result attained is limited to the dataset used and may not be generalized for the population.

4.2 Measure Phase

The first phase of the DMAIC methodology, according to Laux (2017), is the measure phase. For the purpose of this research study, on the basis of information about the measure phase provided by Laux et al. (2017), a data collection plan was generated in this phase wherein, along with the

plan, the data sources and description of the dataset used was mentioned; along with this, the basic performance measures were also mentioned.

4.2.1 Data Collection Plan

One of the assumptions stated by the researcher for this specific research was that the E-commerce organization's primary requirement was assumed by the researcher itself in order to answer the research question. The research question, while dealing with the analysis of online customer reviews, it is specifically with respect to the online clothing/ e-retailing store. Hence the kind of data required for research purposes was an online customer reviews dataset, specifically pertaining to an online shopping portal of a clothing store. The dataset chosen for the purpose of this research was the women's clothing e-commerce reviews dataset published by Brooks (2018).

Since the dataset which was used, has been previously published, it did not include any explicit collection of dataset from the researcher's side; instead, the focus was on using the secondary data made available from a third-party for research purpose in order to focus on answering the research questions considered. This dataset published by Brooks (2018) "consists of reviews written by real customers, hence it has been anonymized" (Agarap and Grafilon, 2018); in the sense, the names of the consumers were not included in the dataset, moreover, references to the organization were also replaced by using another word called retailer.

This women's e-clothing store's dataset consists of 23486 rows, wherein every row represents a review of a single customer. It consisted of 10 feature variables corresponding to each of the customers' reviews. These feature variables include (Brooks, 2018):

- Clothing ID:
    - It is a unique ID number attached to different products/clothing items. The length or total count of such unique clothing IDs, which were integer variables, was

calculated and found to be 1206 for the dataset. The item ID contained within this field can help indicate the clothing item that the respective customer is talking about.

- Age:
    - This next field is the column called Age, which was an integer, was primarily containing the age of the reviewer. The length or total count of unique age categories was calculated and found to be 77. This means that customers who have given reviews on this women's clothing site are found to fall in 77 different age groups.
- Title:
    - This field contained a title given to the review; it was in string format. In comparison to the review itself, the title field in this dataset was more concise and shorter in length. It is a short description of what the full review text consists of. The total count of such unique titles was calculated and found to be 13994.
- Review Text:
    - This field consisted of the review itself, in string format, that was provided by the respective reviewers. The total count of these reviews, in text format, was calculated and found to be 22635 for this dataset.
- Rating:
    - This field consists of an integer variable depicting the rating provided by the reviewer for the specific product their review was based on. The reviewer had the option of giving a product a score from 1 to 5, wherein, 1 would account for worst, and 5 would account for best.

- Recommended IND:
    - The next field in the dataset depicts whether the reviewer recommended a specific product or not. It is a binary variable, meaning, if the reviewer recommended the product, it was given a value 1, else 0.
- Positive Feedback Count:
    - The next variable was a positive integer stating the number of customers, other than the reviewer who was giving that specific review), who found that specific review to be of help.
- Division Name:
    - This field, which was categorical in nature, was to denote the specific division under which the product, which was reviewed by the reviewer, falls. The total count of the number of divisions that a product could fall under was calculated and found to be 3.
- Department Name:
    - This field, which was categorical in nature, was to denote the specific department under which the product, which was reviewed by the reviewer, falls. The total count of the number of departments that a product could fall under or be a part of was calculated and found to be 6.
- Class Name:
    - This field, which was categorical in nature, was to denote the specific class under which the product, which was reviewed by the reviewer, falls. The total count of the number of classes that a product could fall under was calculated and found to be 20.

4.2.2 Basic Performance Measures

The basic performance measures for the purpose of this research were assumed by the researcher. The goal for the organization is: to achieve revenue growth; customer repurchase intention, which would lead to customer retention; and, increase in customer base. While there were multiple strategies that could be applied while aiming for growth, the key performance indicator (KPI) that was considered for this project was customer satisfaction; this KPI identified would help towards achieving the goals stated.

According to Durkacova (2012), KPI can "reflect strategic performance and success of a company" (p. 1080), hence the researcher identified and considered one specific way to be able to gauge the strategic performance of the e-commerce organization, which was by performing sentiment analysis, and by further generating a regression model, and performing topic modeling. The quantitative indicators that would assist in achieving the KPI is as follows:

- Increase in number of positive sentiment reviews
- Decrease in number of negative sentiment reviews
- Decrease in the number of issues faced by customers

4.3 Analysis Phase

The next phase in the DMAIC methodology is the Analysis phase; this is an exceptionally important phase for the framework proposed by Laux et al. (2017) since the authors proposed the inclusion of the big data analytics techniques in: a.) Measure Phase – By generation of analytical dataset containing either historical dataset or data attained by observation and so on, as stated in the Measure Phase above; b.) Analysis Phase – By cleaning and preprocessing the dataset, and then performing necessary analysis tools and techniques. Laux et al. (2017), in their framework, proposed using R programming for performing in-depth analysis; taking this into account, the

researcher, for this research, made use of R for applying big data techniques on the dataset used, and thereby found the solutions for the research questions asked. The programming language R, according to Laux et al. (2017), is one of the "very accessible software programs to process data collected" (p. 671). Many researchers have previously used R tool popularly in their respective areas because of the various advantages offered by this specific programming language such as it is open-source and free – this would have a positive impact on the e-commerce organization's cash flow analysis of the project; its reproducibility in terms of analysis being performed – meaning, any analytical procedures opted by the specific team within the e-commerce organization, and code is written can be shared, and solutions can be reproduced as required; next, R is instrumental in generating "high-quality graphics" (p. 286) – meaning, the reports to be generated and shared with the upper-level management of the e-commerce organization can include various graphs and charts of the results, thereby helping the project team "to produce more effective data visualizations" (p. 286) which can clearly help understand the analysis being done (Mizumoto and Plonsky, 2015).

The deliverables of this phase have been explained by the researcher within each sub-part/section included below.

4.3.1 Data Preparation

The notion of starting the work of analyzing a dataset begins with the very first step after data collection – which includes preparing "quality data by pre-processing the raw data" (p. 375) (Zhang et al., 2003). Many researchers have found this step to be extremely crucial in their research work; the whole idea being that if various tools and techniques of analysis are applied directly to the data which is collected, the results attained may not be accurate if (Zhang et al., 2003):

1. The raw data is inconsistent – meaning, it contains discrepancies in the names or values of a column or document
2. Data is noisy in nature – meaning, it has errors existing
3. It may not contain all the required attributes needed for analysis purposes, and might be incomplete, and so on.

For these reasons, the researcher, by taking guidance of the analysis phase steps from the proposed framework by Laux et al. (2017), prepared] the data by cleaning and pre-processed it for further usage. The following steps were taken by the researcher for performing data preparation in R, which included (R Core Team, 2019):

- Checking the format in which raw data is stored. Since the dataset used by the research in this research work was originally stored in structured CSV format, as published by Brooks (2018), the researcher just read this CSV datasheet into R by making use of the read.csv() function (R Core Team, 2019). If the originally collected raw data was not stored in the CSV format, the researcher would have converted the datasheet in this format, and then further made use of the respective function in order to convert to the right format.
- Next, the researcher made use of the rename() function by R Core Team (2019) in order to assign column names which were better to understand, and easier to refer to whenever required. Next, the researcher checked the dataset read into the R script in order to see if there are any seen discrepancies or errors in format or names, and so on. To understand the columns and its content a little better, the researcher counted the unique values existing in each of the columns, which were also stated by the researcher in the data collection plan section.

- Upon generating the unique values of each of the columns, it was discovered that the column called Division, which contains a higher level division name under which a product may fall, contained 3 unique values – General, General Petite, Initmates (Brooks, 2018). It looked like the division name, Initmates, contained a spelling error, and so the researcher changed this division name to Intimates so that the meaning of the word is correctly understood while referencing it.
- Further, the researcher checked for missing values in the columns they were going to study further during analysis. For this, the researcher chose to check the following columns to understand if there are any values: Review, Rating, Age, Recommend.
- If there were any missing values found, the plan was to remove that specific row that had missing values in any of the column names that the researcher focused on. Upon running the required function, it was found that none of the rows in the dataset have missing values, both NA and NaN, for either of the column names that the researcher chose.
- After checking for missing values in the specific columns, and then deleting the respective rows where a blank value was found – the total number of rows in the dataset was reduced to 22628 reviews. The unique distinct values were calculated for each of the columns, and were found to be as following:

  Clothing_id → 1172, Age → 77, Title → 13984, Review → 22621, Rating → 5, Recommend → 2, Feedback → 82, Division → 3, Department → 6, Class → 20
- Furthermore, the researcher also went through the dataset to see if there any garbage values existing; also, if there any spellings/words that need to be changed in major unique department names, division names, and, class names.

The above-stated steps were all the steps taken by the researcher in the initial data preparation phase, primarily called cleaning the dataset. For further understanding of the overall internal structure of the object of the review, the researcher used the str(reviews) function. This helped with understanding the class of the object of the review, along with get datatypes for the different fields existing within the dataframe.

Further, the researcher then performed steps for preprocessing the text within the dataset, which was the next step in the data preparation phase that the researcher worked on. Many scholars and researchers like Denny and Spirling (2018) have, in their previous research works, stated the importance of performing preprocessing on the textual data existing their dataset for research purpose, with the intention of making "inputs to a given analysis complex" (p. 168) in an efficient way such that the interpretability does not get affected (Denny and Spirling, 2018). Amongst the various approaches and steps within the bag of words representation, as stated by Denny and Spirling (2018), the researcher for this research work performed the text preprocessing steps using the tm package in R, which was published by Feinerer et al. (2008); these steps include:

- First, the researcher generated a corpus of the text data, specifically of the data within the Review column, since the way to documents could be managed by using the tm package is by generating a corpus. By implementing this step, the researcher created a corpus from a character vector of the column Review from the original dataset, thereby converting the Review field to a format on which various functions from the tm package can be applied so as to preprocess the text.
- Next, all punctuations and special characters were removes from the values in the Review field, where the primary thought process was that the inclusion of punctuation marks in

this dataset would not be considered informative. For this, the tm_map() function of the tm package was made use of by the researcher; and the same was used to perform all the other preprocessing steps mentioned in the points below.

- Further, all the numbers from the text contained within the Review field were also removed. Since the focus was going to be on the text contained within the Review field, the researcher removed all the numbers contained within.
- All of the letters of the words in the text were then converted to lowercase since the researcher found that keeping the letter in uppercase or lowercase would not have an impact on what that word means in this case, so it would not feel consistent to keep both uppercase and lowercase letter in words "as two separate word types for the sake of corpus analysis" (p. 171) (Denny and Spirling, 2018).
- Next, the researcher performed stopword removal, in order to get rid of words which "are unlikely to convey much information" (p. 171) (Denny and Spirling, 2018). For this, the stopwords("English") function was used, to remove the stop words which are a part of the default character list that gets generated.
- Furthermore, stemming operation was performed for "reducing a word to its most basic form" (p. 171) with the primary intention of denoting all of the forms of a single word to the simplest basic form, thereby reducing the vocab (Denny and Spirling, 2018).

These were the steps taken by the researcher for applying transformation mapping to the corpora generated.

4.3.2 Exploratory Data Analysis

The researcher conducted Exploratory Data Analysis (EDA) for exploring and detecting different patterns in data empirically, especially since the study was deemed to be exploratory (Jebb et al.,

2017). In their research, the authors Jebb et al. (2017) have stated how there have been very few research studies on EDA, and went on to argue that EDA has multiple benefits attached to it and "it maximizes the value of the data" (p. 266); also the authors have stated that involvement of EDA is exceptionally important to make fullest use of the data available. So, by making use of this type of analysis on the text available within the women's clothing e-commerce dataset by Brooks (2018), the researcher understood and documented the basic structure of the contents of this dataset.

There were various steps which included the tasks performed to explore the data; these included:

- Many authors, such as Gangolly and Wu (2002), have stated how generation of term-document frequencies (TDM) to understand how frequency of terms in documents within the corpus can help realize its importance. The researcher generated a term document matrix of the corpus, with the primary intention of converting the generated TDM to a matrix class format, and to finally get the sorted list of frequencies of terms to understand which words have customers mostly focused on in their respective online reviews.
- The researcher used the TermDocumentMatrix() function to generate a TDM of the corpus, and further generated a matrix out of it, based on which terms with top frequencies were displayed.
- The top 20 most frequently used words in the whole corpus of the Review field were as shown in Figure. First, these top 20 terms were generated, as suggested by Vivek (2018), to understand that among the 23486 records of customer reviews, which are the 20 terms that have been used the most by the customers. From the Figure, it can be seen how often

the words the top 5 words have been used even in comparison to the other terms; these top terms in the dataset were: dress, love, fit, size, and, look.

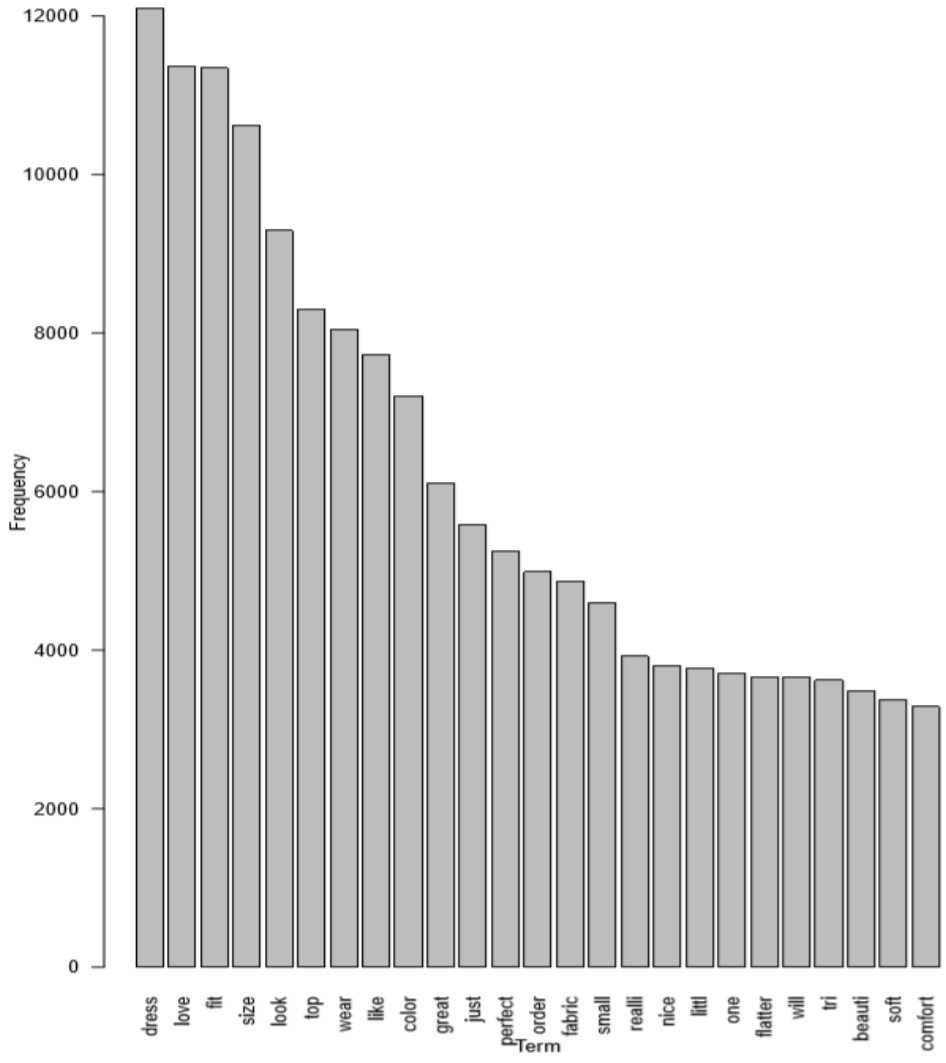


Figure 4.1 Top 20 terms bar chart

- Research studies have previously used word clouds in opinion mining projects and projects where feedback of consumers is involved; this is because “deeper insights about actual review content are needed for users to make better decisions” (p. 151) (Wang et al., 2014). In order to summarize the contents of the online reviews’ corpus visually, the researcher

generated word clouds showing words in different colors and fonts size “mapped to the word frequency, popularity or importance” (p. 151) (Wang et al., 2014).

- The first word cloud generated by making use of the code published by Vivek (2018) was done to be able to visually see the top 100 most used words in the dataset, which is as shown in the Figure. With the help of this cloud, the most used words stand out clearly because of their font size, thereby showing its importance as stated by Wang et al. (2014).

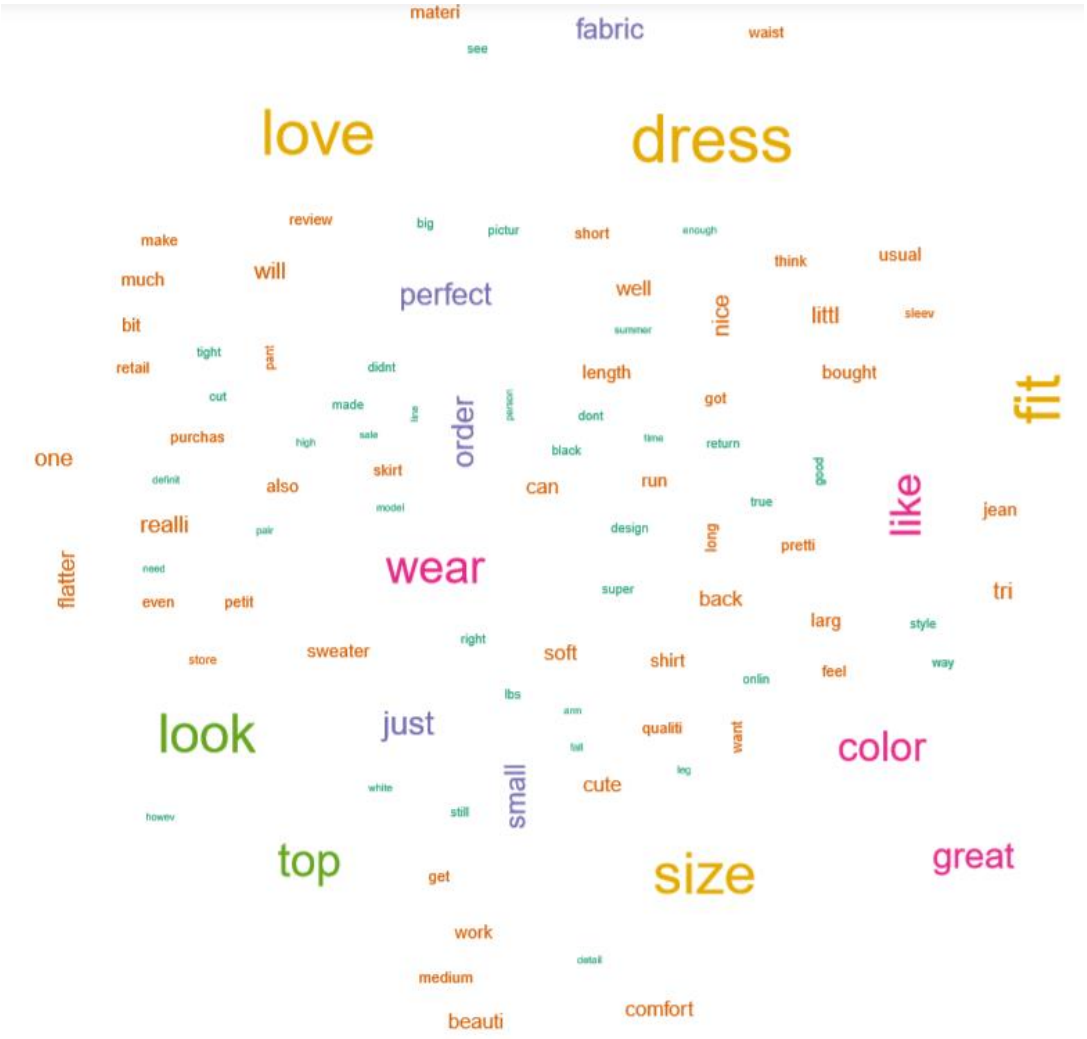


Figure 4.2 Top 100 terms word cloud

- Next, by looking at the dataset along with its field names and contents, a column called Recommend consisted of values 1 or 0 based on whether the customer of the specific review recommended that specific product on which the review was written. The dataset was split into two different data frames – one consisting of reviews of customers who recommended the product, and the other of the reviews of customers who did not recommend the product.
- For the users who wrote the review and recommended the product, the researcher wanted to understand the most used words by such customers, as suggested by Vivek (2018), in order to see what these customers gave importance to. To do this, from the data frame, a corpus was generated of reviews of users who recommended the specific product; after which, with the help of TDM generated, the top 20 most used words were visually displayed as shown in the Figure.

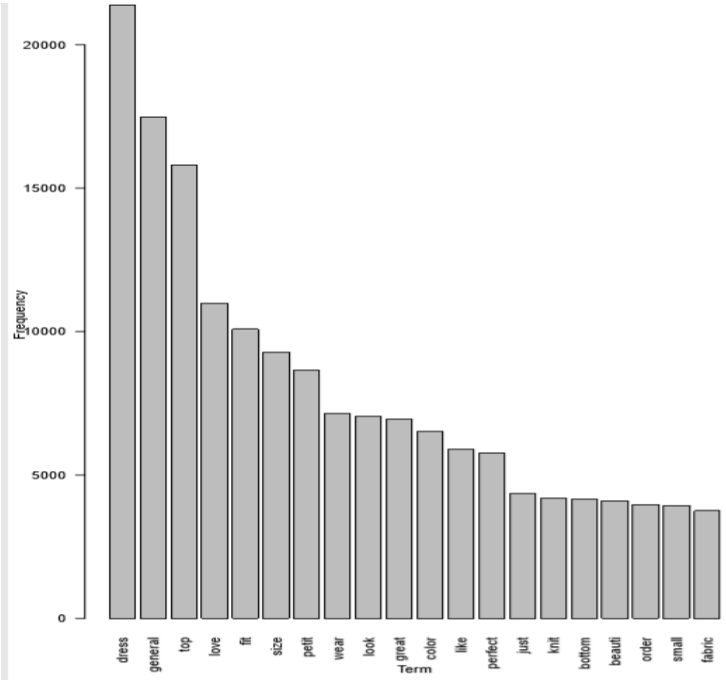


Figure 4.3 Top 20 terms of customers who recommended the product

- Next, for the users who wrote the review and did not recommend the product, the researcher wanted to understand the most used words by such customers, in order to see what these customers gave importance to. To do this, from the data frame, a corpus was generated of reviews of users who did not recommend the specific product; after which, with the help of TDM generated, the top 20 most used words were visually displayed as shown in the Figure.

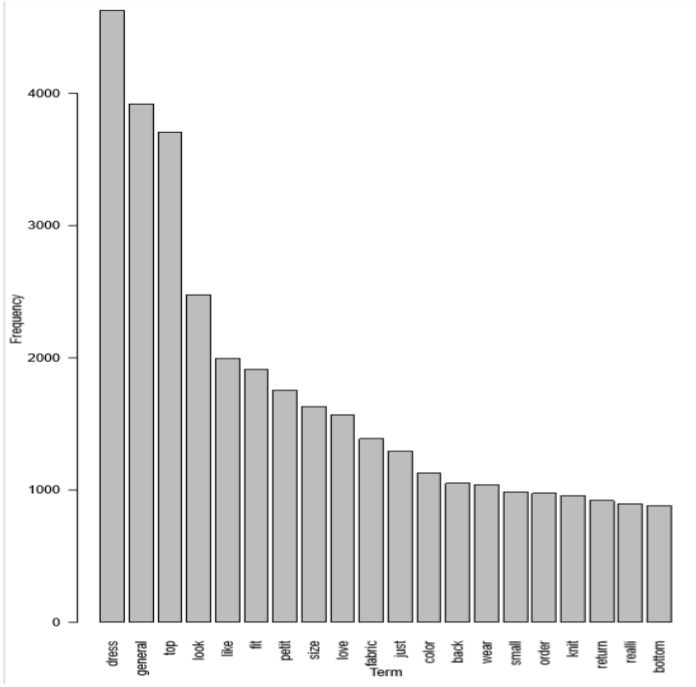


Figure 4.4 Top 20 terms of customers who did not recommend the product

- Just the initial look at the bar graphs, differences in the usage of words could be noted – with words such as back and return being a part of the corpus consisting of reviews where the product was not recommended; as opposed to the words such as great and perfect appearing in the corpus consisting of reviews where the product was recommended.
- In order to further understand the difference in reviews of customers who have recommended the product versus the ones who have not recommended, specifically in

terms of the usage of words, a comparison cloud was generated from the wordcloud package by Fellows (2018). For the purpose of this research, the researcher compared and contrasted between the two corpora to see the terms which have been used specifically only in the reviews of users who have recommended to the product, as well as, the terms which have been used specifically only in the reviews of users who have not recommended to product side by side in a word cloud, the researcher generated a comparison cloud as suggested by Vivek (2018), by making use of the wordcloud package by Fellows (2018), as shown in the Figure. From the Figure, the users who did not recommend the product had terms such as cheap, return, disappointed, and so on.

Figure 4.5 Comparison cloud

To explore which department have the users focused on in their reviews, the next step by the researcher consisted of generating a percentage of reviews for each department within each division. The reason for choosing to understand the distribution of reviews

department-wise was to see which department has gained the most attention since the percentage of the total reviews in that department. Based on the total percent of reviews generated department-wise, the Department Tops received the greatest number of customer reviews, followed by Dresses, Bottoms, and so on. The percentage of reviews within each division were as follows - General: 59.1%, General Petite: 34.6%, Intimates: 6.3%; and, percentage of reviews by Department were as follows – Tops: 44.4%, Dresses: 27.2%, Bottoms: 16.2%, Intimate: 7.3%, Jackets: 4.4%, Trend: 0.5% , as shown below in Figure 4.6.

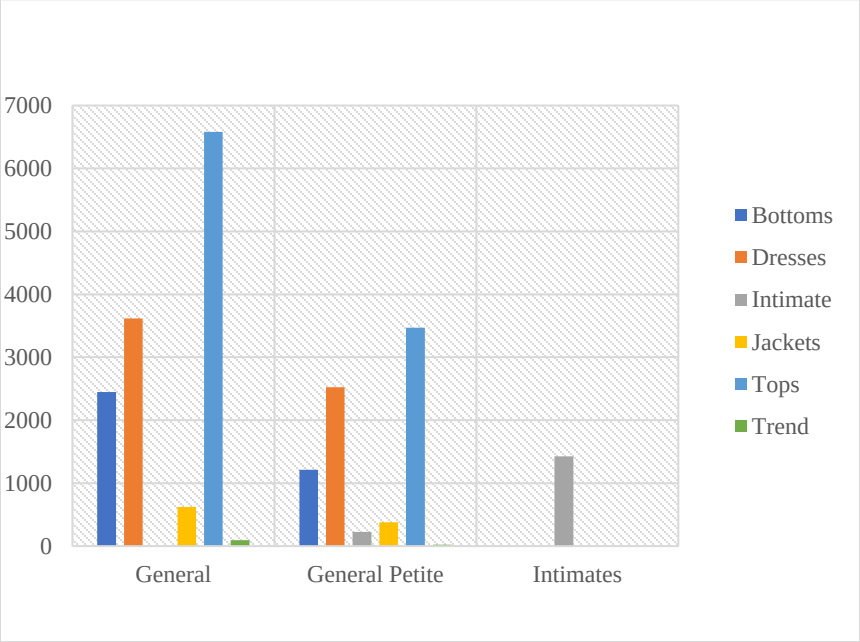


Figure 4.6 Reviews by department by division

- Now that the distribution and spread of reviews of all of the customers were understood department-wise, the researcher then went on to explore the percentage of the reviews

department-wise of users who have recommended the product, as well as, the ones who have not recommended the product. Observations were as mentioned below:

- For the reviews of customers who recommended the product, the percentage of reviews department-wise was as shown in the Figure. The total number of such reviews where products were recommended was found to be 18527.

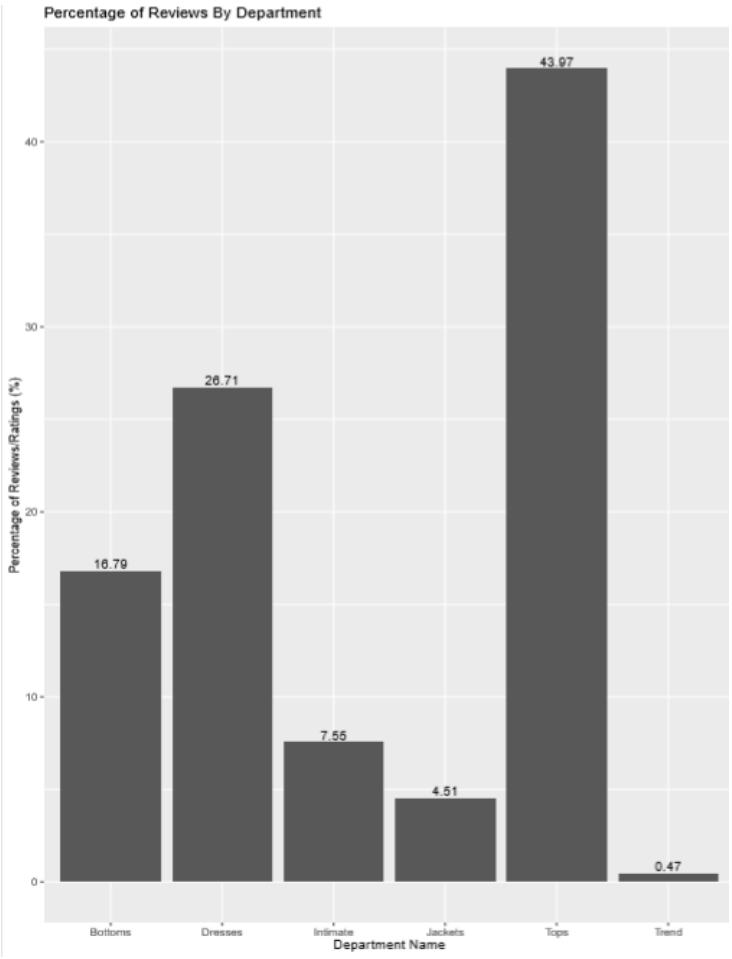


Figure 4.7 Percentage of reviews by department (Product recommended)

- For the reviews of customers who did not recommend the product, the percentage of reviews department-wise was as shown in the Figure. The total number of such reviews where products were not recommended was found to be just 4101.

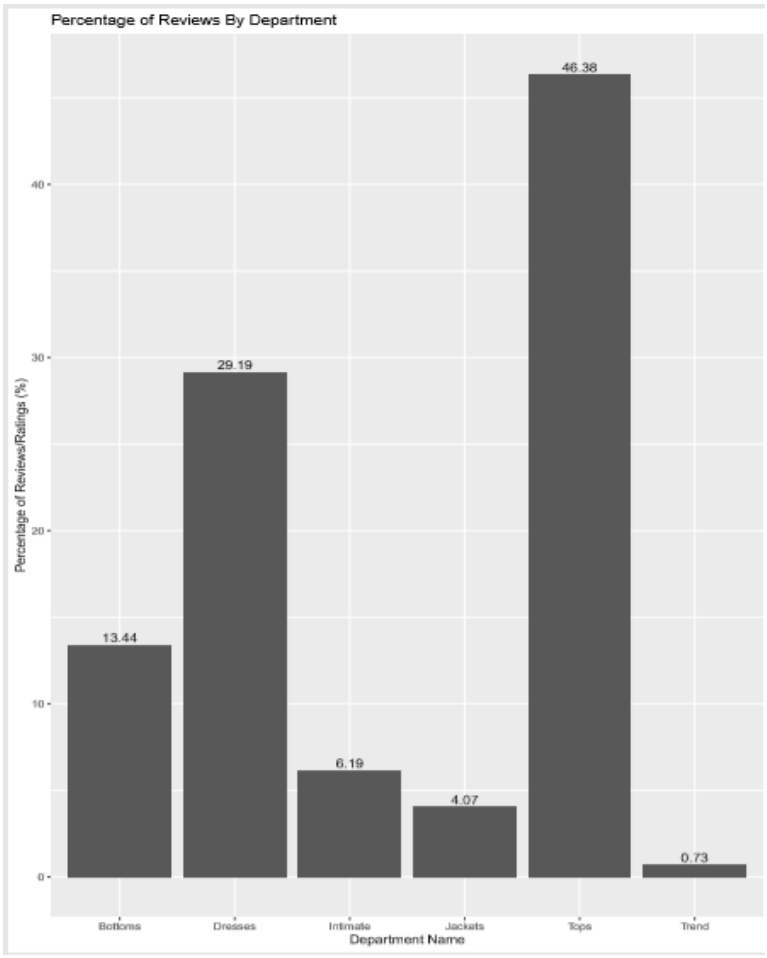


Figure 4.8 Percentage of reviews by department (Product not recommended)

- Exploring the Age field in the dataset published by Brooks (2018) could help the organization understand its users by realizing the age structure of the population getting attracted to their online shopping portal. For this, customers under the age of 30 were put in the 18 to 29 age group bracket; customers under the age of 40 but over 29 – were put in the 30 to 39 age group bracket; customers under the age of 50 but over 39 - were put in the 40 to 49 age group bracket; customers under the age of 60 but over 49 - were put in the 50 to 59 age group bracket; customers under the age of 70 but over 59 - were put in the 60 to 69 age group bracket; customers under the age of 80 but over 69 - were put in the 70 to 79

age group bracket; customers under the age of 90 but over 79 - were put in the 80 to 89 age group bracket; and finally, customers under the age of 100 but over 89 - were put in the 90 to 99 age group. In this case, the customers who had given most reviews were roughly between the age groups of 30 years to 50 years. Also, as seen in Figure 4.9, after an initial rise in the number of reviews acquired from customers in the 18-29 years age group to 30-39 years age group, the reviews seemed to have gradually decreased from the customers in the 40-49 years age group and further.

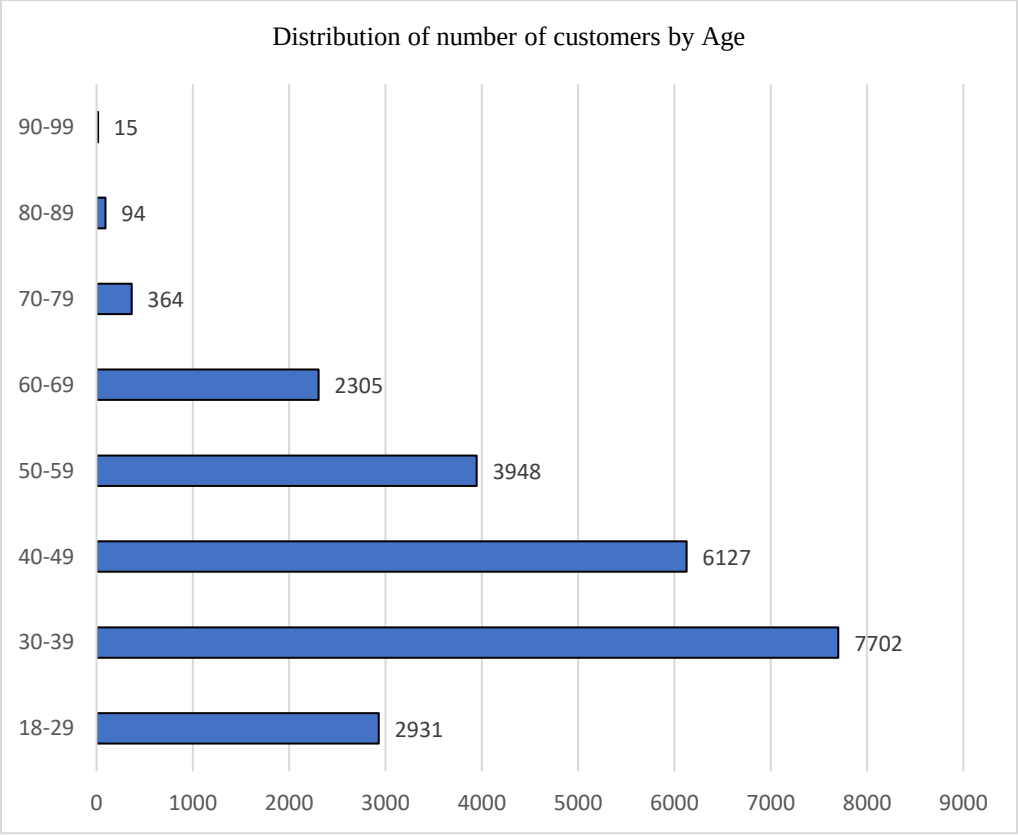


Figure 4.9 Total reviews by customers by age

- Further, to dive deeper, the number of reviews by department by age, for the dataset published by Brooks (2018), were calculated and have been shown in the Figure 4.10

below. This helps the organization understand which products were preferred, or have been reviewed most, by customers of specific age group.

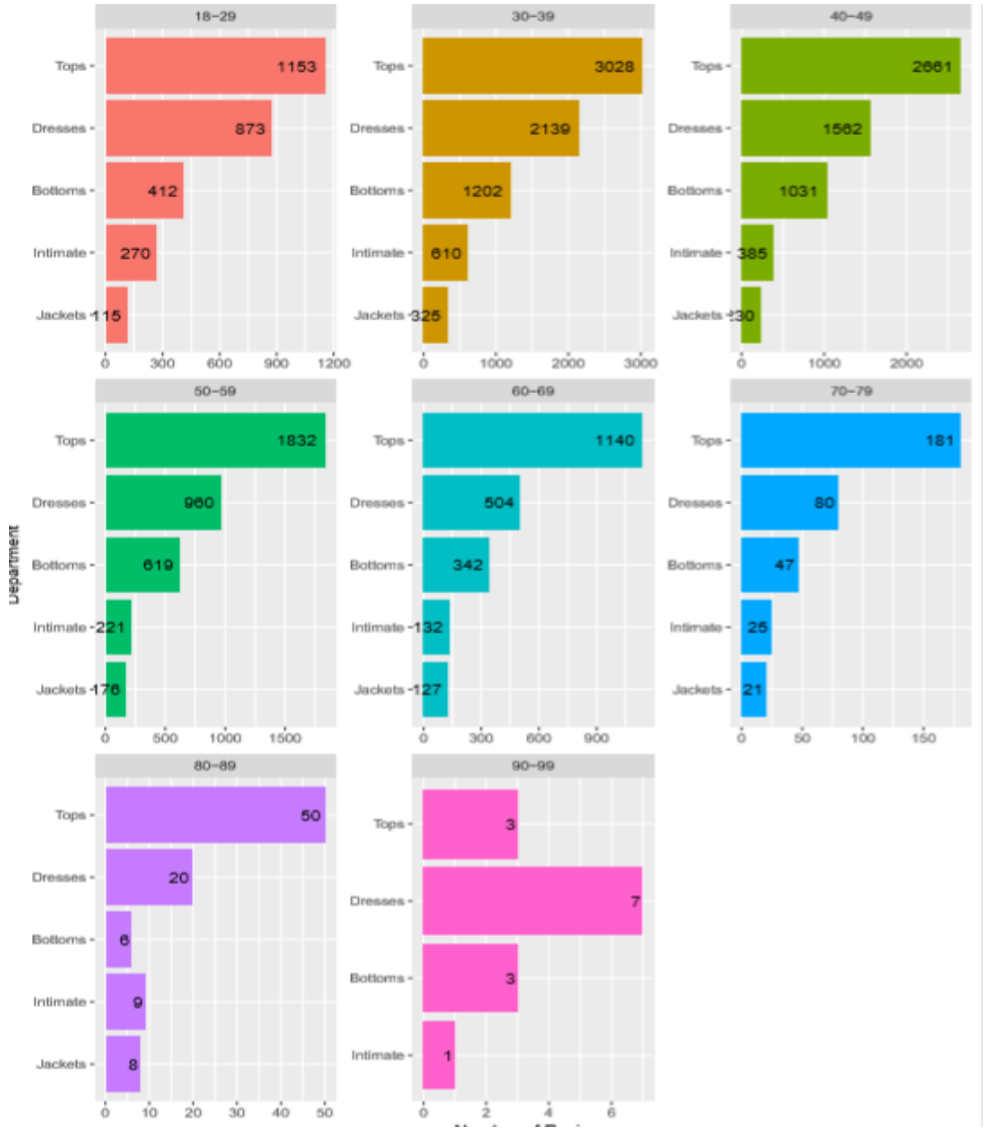


Figure 4.10 Number of reviews by department by age

- Sentiment Analysis -
    - The next step in this process of exploring the dataset was to understand the opinion of the customers who had given reviews on the online portal. The most basic reason

why and how this analysis would help the e-commerce organization was by mining the users' thought process or preference and analyzing which products and/ factors of the online portal could be improved upon. To mine the opinions, in other words, to perform opinion mining, the technique called sentiment analysis exists; this can help by recognizing the emotions within the text included in the dataset by stating its polarity and/ or "by rating how positive or negative it is on a scale" (p. 2) (Veiga, 2019). As stated by Bagheri et al. (2013), many researchers have made use of sentiment analysis (or opinion mining) primarily for projects or research work done within natural language processing area and/ data mining related area, especially for analyzing customer reviews. While multiple approaches and algorithms can help with performing sentiment analysis, the kind of result one wants to acquire would determine the technique one would have to follow; according to Veiga (2019), different approaches could be taken, these involve lexicon-based approach, and, text classification approach.

- For this research, the researcher made use of tidytext package by Silge and Robinson (2016) in R for performing sentiment analysis using AFINN lexicon. This AFINN lexicon was published by Nielsen (2011). The tidytext package consists of many such "sentiment lexicons in the sentiments dataset" (p. 14); this dataset comes along with the package (Silge and Robinson, 2017).
- The primary reasoning behind choosing to make use of AFINN, as opposed to other lexicons, was because the way this lexicon works is by assigning a score between -5 to +5 to words, thereby also categorizing them as positive (words which get positive scores assigned) and negative (words which get negatives scores assigned)

(Silge and Robinson, 2017). The main advantage is that with this lexicon, a score will be assigned to every review of the customer review dataset, which can then be used further by the e-commerce organization for deep diving into analyzing the impact and correlation of other factors with the sentiments of the customers, as opposed to using other lexicons such as BING which would essentially just be used to state if a review was positive or negative, and so on (Silge and Robinson, 2017).

- The researcher used the get_sentiments() function for specifically getting the AFINN lexicon; Review.Text field, of the dataframe that was used, was tokenized, stop words and other punctuations were removed. Once the afinn lexicon table was invoked, it was inner joined with the customer reviews dataset; the data frame thus generated consisted of the sentiment score for each review, which was the mean of afinn scores of words in that respective review, along with other fields. According to the summary table and graph generated, by taking the code published by Michel (2019), and using the Rating field and mean afinn score (sentiment score) as shown below, there seems to be a correlation between the two – which can certainly help the organization to look into the relationships of these two fields to gather more insights. The reference for the code for performing sentiment analysis using afinn was taken from Littlejohn (2018).

Table 4.2 Mean and sd of sentiment score by rating

	Rating	mean	sd
1:	1	0.5053558	1.3976720
2:	2	0.8685473	1.2831571
3:	3	1.2153937	1.1618703
4:	4	1.6049195	1.0216687
5:	5	1.9056631	0.9120086

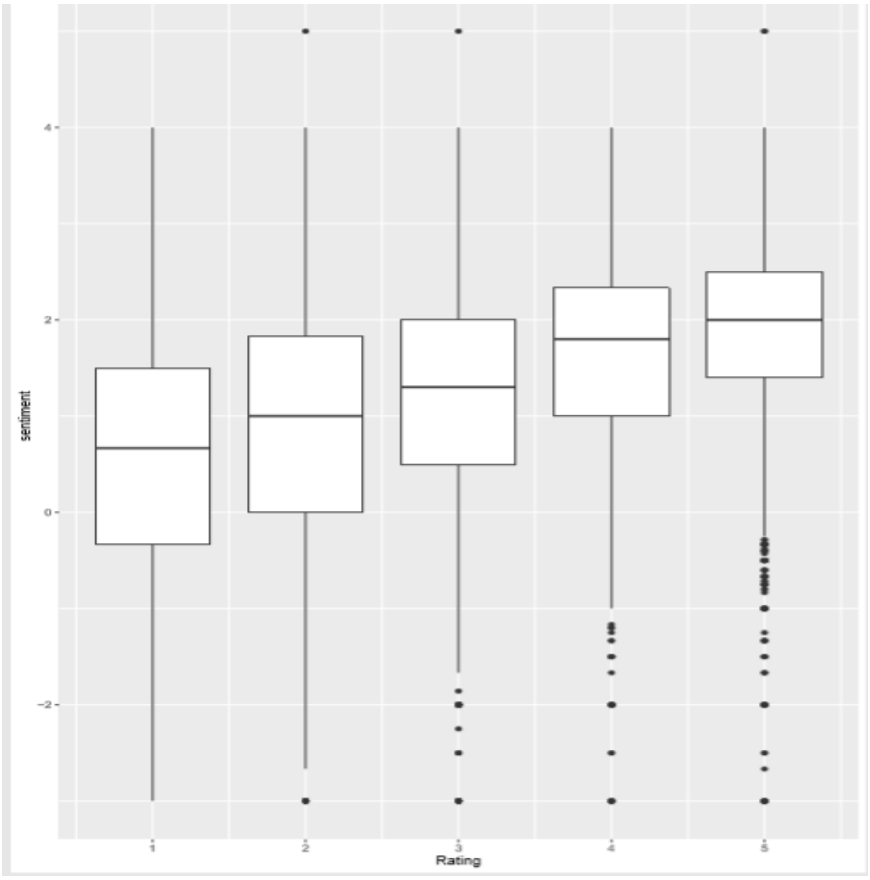


Figure 4.11 Box plot showing ratings by sentiment score

- While the AFINN lexicon used did show a correlation that exists in the respective e-commerce organization's dataset, it was not extremely significant for this data. Based on the AFINN lexicon word list that was used for giving a score to the words and generating an afinn or sentiment score, it was found that out of the 21855 reviews which had words in the AFINN dictionary with scores attached to them:
  - Total number of reviews in the dataset characterized as positive reviews (having afinn mean score above 0) were: 19875 (91%)
  - Total number of reviews in the dataset characterized as negative reviews (having afinn mean score below 0) were: 1366 (6%)

- Total number of reviews in the dataset characterized as neutral reviews (having afinn mean score equal to 0) were: 614 (3%)

- This above observation can be used by the e-commerce organization to understand the trend in their dataset based on the polarity of the reviews within their customer reviews dataset. In this case, mean afinn score which gave the sentiment score to respective reviews, almost 91% of the reviews were found to have positive polarity, whereas, only about 6% of the reviews were found to have negative polarity, and about 3% of the reviews were found to have neutral polarity.
- From the boxplot generated in Figure 4.11, it can be seen that certain outliers are existing, so the next step, as stated by Littlejohn (2018), was to explore these outliers to get some insight about the types of reviews which act as outliers in the given dataset:
    - The first step conducted was to generate a summary of the data frame created after adding the sentiment column to it which contained sentiment score (mean of afinn score) for the respective review. The summary of the columns of the table was as shown in the Figure 4.3 and Figure 4.4.

Table 4.3 Summary of ReviewID, and Age

ReviewID	Age
Min. : 1	Min. :18.00
1st Qu.: 5674	1st Qu.:34.00
Median :11338	Median :41.00
Mean :11327	Mean :43.28
3rd Qu.:16987	3rd Qu.:52.00
Max. :22628	Max. :99.00

Table 4.4 Summary of Rating, Recommend, and sentiment

Rating	Recommended.IND	sentiment
Min. :1.000	0: 3870	Min. :-3.000
1st Qu.:4.000	1:17985	1st Qu.: 1.000
Median :5.000		Median : 1.875
Mean :4.198		Mean : 1.638
3rd Qu.:5.000		3rd Qu.: 2.400
Max. :5.000		Max. : 5.000

- Creating a summary table helps the organization gauge the dataset further for figuring out the outliers that may exist in the data collected. From the summary table generated the sentiment scores given to reviews in the dataset range from -3 to +5 value.
- The researcher then examined the reviews which have a very high sentiment score attached to it but have been given a low rating of 1 or 2. Of the 21855 reviews in the dataset, it was found that 114 reviews have a sentiment score which is greater than or equal to +3.00, but a rating of just 1 or 2 given to them, and can, therefore, be considered as outliers. To get this information form the data frame, the following code pertaining to the filter() function was written: filter((Rating == 1 | Rating==2) & sentiment >= 3.00).
- Next, the reviews which had a very low sentiment score attached to it but have been given a low rating of 4 or 5 were examined. Of the 21855 reviews in the dataset, it was found that 29 reviews have a sentiment score which is less than or equal to -3.00, but a rating of just 4 or 5 given to them, and can, therefore, be considered as outliers. To get this information form the data frame, the following code pertaining to the filter() function was written: filter((Rating == 5 | Rating==4) & sentiment <= -3.00).

- Performing this step for analysis purposes can help the organization see not only if there are outliers existing, as suggested by Littlejohn (2018), but to then dive into the aspects of why certain reviews have a very high rating but a very low sentiment score, and vice versa.

- Linear Regression:
  - Authors such as Fox (2003) have made use of regression by generating linear models for understanding the impact of certain dependent variables on an independent variable. Looking at the dataset considered for this research, the next step taken by the researcher to understand the impact of different variables such as Age, Recommend, Department, and Rating on the independent variable (sentiment score). This step helps the e-commerce organization by stating which field within the dataset should they focus on the most, because of the highest impact that it has on the sentiment score.
  - The researcher generated a simple regression model, as shown by Fox (2003), by taking the y variable as the sentiment, and look at the impact of the Age, Rating, Recommend, and Department fields on the sentiment score.

```
Call:
lm(formula = new_reviews$Sentiment ~ new_reviews$Age + new_reviews$Rating +
    new_reviews$Recommendation + new_reviews$Department)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9239 -0.5867  0.0892  0.6776  4.1555

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.2022977  0.0404513   5.001 5.75e-07 ***
new_reviews$Age             -0.0001360  0.0005596  -0.243 0.807976
new_reviews$Rating           0.2875147  0.0101386  28.358  < 2e-16 ***
new_reviews$Recommendation1  0.2153379  0.0293820   7.329 2.40e-13 ***
new_reviews$DepartmentDresses  0.0649828  0.0215769   3.012 0.002601 **
new_reviews$DepartmentIntimate  0.0731858  0.0307505   2.380 0.017322 *
new_reviews$DepartmentJackets  0.0574538  0.0369823   1.554 0.120307
new_reviews$DepartmentTops   0.0721650  0.0199229   3.622 0.000293 ***
new_reviews$DepartmentTrend -0.0998542  0.0983862  -1.015 0.310155
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.016 on 21846 degrees of freedom
Multiple R-squared:  0.1266,    Adjusted R-squared:  0.1263
F-statistic: 395.9 on 8 and 21846 DF,  p-value: < 2.2e-16
```

Figure 4.12 Screenshot of output of linear regression model

- The output of running the simple regression model was as shown in the above Figure. From the output of running the linear regression model, the following can be interpreted:
    - The estimated value of the Age field is negative (-0.001); this means that as the value in the Age field increases, the sentiment score decreases, and vice versa. This means that, with respect to the dataset used in this research, the older aged customers are not happy, or may not have used extremely positive words in their reviews; however, this is not statistically significant in nature.
    - The estimated value of the Rating field is positive (0.2875); this means that as the value in the Rating field increases, the sentiment score also increases. This means that, with respect to the dataset used in this research, rating given by the customer has a high impact on the sentiment score – especially since it is statistically significant at 1% level.
    - The estimated value of the Recommend field is positive (0.21); this means that as the value in the Recommend field increases, the sentiment score also increases. This

means that, with respect to the dataset used in this research, whether a customer has recommended a product has a high impact on the sentiment score – especially since it is statistically significant at 1% level.

- The categorical variable Department was also considered as one of the x variables, in order to find its impact on the sentiment score, as follows:
    - The estimated value of the Dresses Department is positive (0.064). This means that, with respect to the dataset used in this research, the Dresses department does have an impact on the sentiment score – especially since it is statistically significant at 5% level.
    - The estimated value of the Intimate Department is positive (0.073) – slightly higher than the Dresses department. This means that, with respect to the dataset used in this research, the products within this Intimate department category does have an impact on the sentiment score – especially since it is statistically significant at 10% level, so the impact of this Intimate department is less in comparison to the Dresses department.
    - The estimated value of the Jackets Department is positive (0.057). However, it is not statistically significant.
    - The estimated value of the Trend Department is positive (-0.09). This means that, as the value in this department increases, the sentiment score decreases, and vice versa. With respect to the dataset used in this research, the Trend department does not seem to have an impact on the sentiment score, especially since it is not statistically significant.

- The estimated value of the Tops Department is positive (0.072). This means that, with respect to the dataset used in this research, the Tops department does have an impact on the sentiment score – especially since it is statistically significant at 1% level – which is very high in comparison to the Dresses and Intimate departments.

- With the help of the results acquired by generating a simple linear regression model, Rating had a high impact on the sentiment scores for this dataset, in comparison to all other fields. This can also further be proved by the results of the ANOVA, which is as shown in the Figure 4.13.

```
> anova(mod)
Analysis of Variance Table

Response: new_reviews$Sentiment
                              Df  Sum Sq Mean Sq   F value    Pr(>F)
new_reviews$Age                1     2.9     2.9    2.8305  0.092506 .
new_reviews$Rating             1  3193.8  3193.8 3094.5822 < 2.2e-16 ***
new_reviews$Recommendation     1    54.9    54.9   53.2050 3.107e-13 ***
new_reviews$Department         5    17.3     3.5    3.3496  0.005014 **
Residuals                  21846 22546.5     1.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4.13 Screenshot of output of anova() run on the linear model

  - This linear regression model helped in understanding factors affecting the customer experience, wherein customer experience is being measured by or defined as the sentiments or opinions of the customers who have written reviews on the online shopping portal of the e-commerce organization.

- Till now, what was understood about the dataset was that the feature variable called Rating had an impact on the sentiment score, and it was statistically significant. This means, the users have given the ratings which match with what they have written in their review. However, to identify the issues within the dataset by analyzing the content of the reviews,

the next step was performing topic modeling using the Latent Dirichlet Allocation (LDA); this had to be performed for inferring topics within the dataset (Kuang et al., 2017). Many researchers such as Calheiros (2017) have made use of topic modeling, more specifically have also used LDA, for the purpose of analyzing the content within their respective datasets for performing text mining and further generating topics. For performing topic modeling, the topicmodels package in R by Grün and Hornik (2011) was used, as shown by Soltoff (2020), in the following way:

- In this case, the researcher did not know of the exact potential topics for the respective e-commerce organization, so the latent topic structure, as mentioned by Soltoff (2020), had to be assumed based on what fits best.
- Out of the 21855 reviews, it was found that the percentage of reviews with ratings 1, 2, 3 were very low in comparison to the percentage of reviews with ratings 4 and 5. The total number of reviews having a 1, 2, or 3 rating was 4912. This means that almost 22.5% of the reviews had been given a lower rating.

Table 4.5 Percentage of reviews by Rating

	Rating	percent
	<int>	<dbl>
1	1	3.49
2	2	6.69
3	3	12.3
4	4	21.6
5	5	55.9

- Since rating had an impact on the overall sentiment, the researcher applied topic modeling by generating a subset of the dataset by filtering for reviews which have been given a low rating of either 1, 2, or 3; the primary reasoning behind applying this filter on the dataset was to look at the reviews of customers who have given a

rather low score or rating to the product they are reviewing. Upon performing topic modeling as shown by Soltoff (2020) on this data, the two topics generated were as in the Figure 4.14 below.

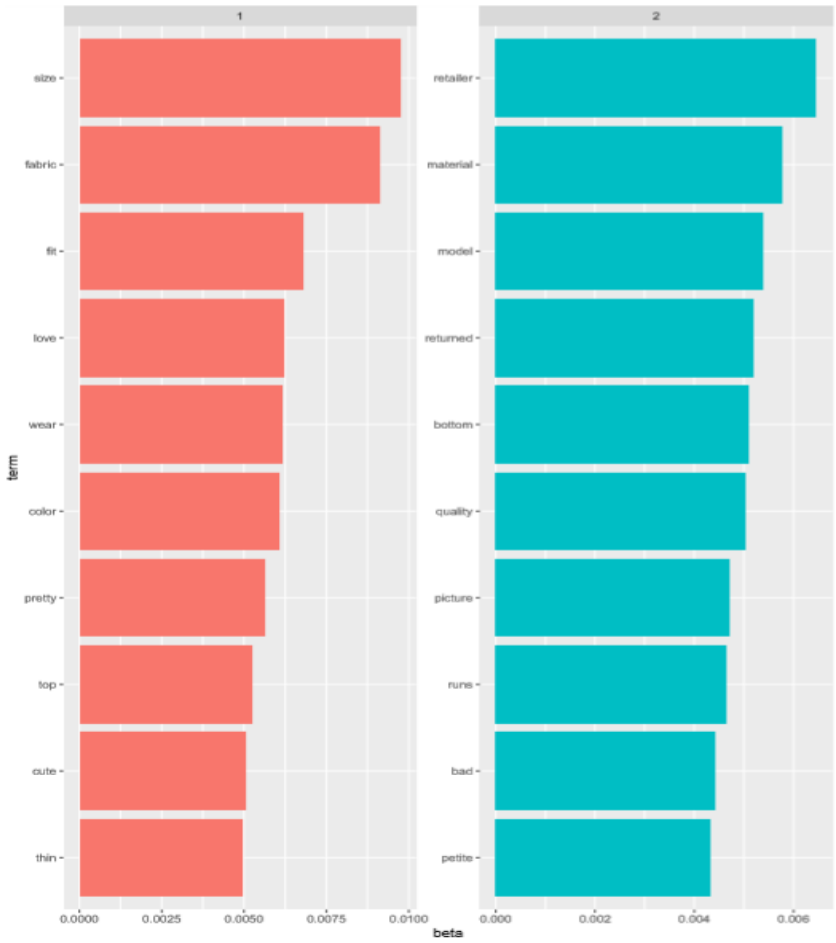


Figure 4.14 Topics of low rating

From the topics generated, a general trend of the first topic seems to be pertaining to fabric and fit, whereas, the general trend of the second topic seems to be pertaining to material and quality – on the basis of the top words appearing in those topics.

- To dig deeper specifically on the reviews which have been given a low rating of 1, 2, or 3, plus these reviews have a negative sentiment attached to them were filtered

for further looking at topics herewith, to point out major issues. Total number of such reviews having negative sentiment score, plus a low rating (1, 2, or, 3) were 791. Topic modeling was performed using LDA, as shown by Soltoff (2020), on this filtered set of reviews, and the two topics generated were as shown in the Figure 4.15 below.

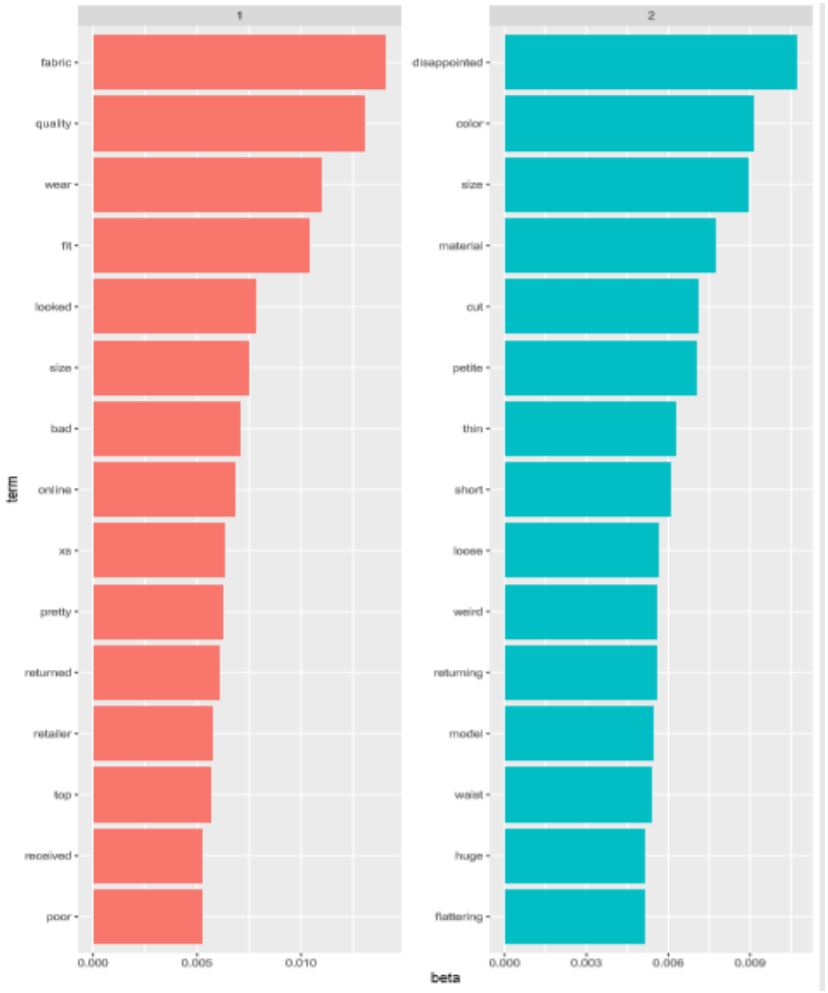


Figure 4.15 Topics identified of reviews with negative sentiment and lower ratings

- Performing this topic modeling, two topics were generated using LDA "by scanning the words and their distribution probabilities" (p. 362) within the subset taken of the dataset with negative sentiment reviews having lower rating (Kuang et al., 2017). The top words

appearing in these two topics were displayed in the Figure, which helps the e-commerce organization focus their direction to the major themes which the customers, who have given a lower rating, have focused on; these are as follows:

  - Fabric
  - Quality
  - Size
  - Fit
  - Material

- LDA does have certain limitations to it, as specified by authors such as Hong and Davison (2010), as well as Moro et al. (2015), these include:
  - Unsupervised LDA may cause an issue when considering texts with more content, especially since it is automated to work by identification of terms for the generation of topics (Hong and Davison, 2010).
  - The number of topics chosen plays a difference in identifying terms that have occurred together; however, the context in which the terms of a topic have been used may not be the same (Moro et al., 2015).
- For the above reasons, to ensure that LDA did generate the right themes/topics, the researcher performed additional in-depth analysis done manually, as follows:
  - A subset of 250 reviews with only negative sentiments were considered
  - By using MS Excel, the researcher went through each of these 250 reviews in order to categorize them into specific topics; this was done based on the wordings within these reviews

- From the categorization performed, the output was a shown in the Figure, which can be described as follows:
    - There was more than one issue observed in many reviews, so the researcher counted each of those issues separately. For the 250 reviews considered, there were total of 361 issues found.
    - The issues observed were of the following categories:
        - Size/Fit – Here, the researcher combined the issues which spoke about the size and/ fit of the products. This issue was most found in the Tops department, followed by Dresses, Bottoms, Jackets, Intimate, and Trend. Since size/fit issues were the most among the reviews, the results attained match with what studies have shown earlier, as stated by Nasibov et al. (2016), that bad fit/size can lead to most dissatisfaction of the customers.
        - Quality - This issue was most found in the Tops department, followed by Dresses, Bottoms, Intimate, Jackets, and Trend.
        - Material – Here, the researcher combined the issues which spoke about the fabric and material of the products. This issue was most found in the Tops department, followed by Dresses, Intimate, Bottom, Jackets, and Trend.
        - Looked different online - Here, the researcher combined the issues which spoke about how the product looked on the online portal, or how the product looked on the model whose photo was on the online portal next to the product, and so on. This issue was most found in

the Tops department, followed by Dresses, Jackets, Bottoms, Trend, and Intimate.

- Price - This issue was most found in the Dresses department, followed by Tops, Bottoms, Jackets, and Intimate.

Table 4.6 Total reviews per issue by department

Row Labels	Count of Review ID
Size / Fit	
Tops	71
Dresses	26
Bottoms	19
Jackets	8
Intimate	7
Trend	3
Quality	
Tops	46
Dresses	17
Bottoms	15
Intimate	10
Jackets	4
Trend	1
Material	
Tops	38
Dresses	8
Intimate	7
Bottoms	6
Jackets	2
Trend	1
Looked different online	
Tops	30
Dresses	10
Jackets	4
Bottoms	4
Trend	1
Intimate	1
Price	
Dresses	7
Tops	6
Bottoms	4
Jackets	3
Intimate	2
Grand Total	361

- The output that was received, by the manual categorization of the reviews taken as a subset, proved that the themes/topics identified by LDA were directionally right - since the issues observed by the researcher using MS Excel and going through each review individually, were the same terms that LDA had provided in the form of the topics generated.

4.4 Improve Phase

The next phase of the DMAIC methodology is the Improve phase wherein improvement plans have been made by taking both existing and potential customers into account as stated by Laux et al. (2017); the researcher also considered the pain points acquired from Analysis phase for suggesting potential strategies that could be implemented for enhancing customer satisfaction. As stated in the Measure phase by the researcher, the KPI for this project has been assumed by the researcher to be an enhancement in customer satisfaction; so the improvement strategies suggested have also been towards the same effort and keeping the same aim in mind for the e-commerce organization. The second research question (RQ2) considered for this thesis was "Which strategies could be implemented to further improve upon factors affecting customer experience?"; the researcher has answered this research question within this phase, in the sections stated below:

- Major issues identified:
    - Size/Fit
    - Quality
    - Material
    - Looked different online
    - Price
- Voice of Customers (VoC) within each issue:
    - Size/Fit
        - The context in which these issues have appeared in some reviews, within the dataset published by Brooks (2018), were as follows:
            - Within Tops Department –
                - "made me look bigger", "huge and shapeless", "bad arms and shoulder fit; super tight in the shoulder area and upper

arms", "super-short", "boxy", "too small compared to other retailer blouses of the same sizes", "baggy fit", and so on.

- Within Dresses Department –
    - "humongous and gave me no shape", "big everywhere except bust area", "looked like a sack", "too tight in strange ways", "loose and baggy through the top area, but tight around thighs"
- Within Bottoms Department –
    - "ran very small; weird fit", "loose in the waist", "horrible pants"
- Within Intimate Department –
    - "oddly shaped", "made me look 20 pounds heavier", "fit was awful",
- Within Jacket Department –
    - "boxy", "sticks out oddly in every direction", "shoulders are bulky"
- Within Trend Department –
    - "extremely oversized"

- Quality
    - The context in which these issues have appeared in some reviews, within the dataset published by Brooks (2018), were as follows:
        - Within Tops Department –

- "thin and poor quality", "rip the first time I wore it", "top tore apart", "it shrunk upon washing according to the instructions", "shrunk significantly, "damaged; had a stain on the front", "feels cheap; not flattering"

- Within Dresses Department –
  - "made very poorly", "two buttons came up before putting it on", "zipper was defective", "looks and feels cheap", "a lot of pieces are missing"
- Within Bottoms Department –
  - "lack of quality", "disappointed in the quality", "quality for the price not there", "poor craftsmanship"
- Within Intimate Department –
  - "product was defective", "required constant readjustment", "not the most comfortable", "outer layer shrunk upon washing"
- Within Jacket Department –
  - "discrepancy between quality and cost", "too delicate"

- Material
  - The context in which these issues have appeared in some reviews, within the dataset published by Brooks (2018), were as follows:
    - Within Tops Department –
      - "scratchy, stiff", "material is cheap and looks torn", "very stiff and heavy", "felt itchy and rough"

        - Within Dresses Department –
            - “material is thin and feels cheap”, “pattern was cut up and looked pieced together”
        - Within Bottoms Department –
            - “fabric was super thin and very cheap; stretchy”, “rough texture”
        - Within Intimate Department –
            - “thin and flimsy”, “itchy”, “lot of static because of polyester”
        - Within Jacket Department –
            - “really bulky and scratchy; feels like wearing an old wool blanket”
        - Within Trend Department –
            - “fabric used made it look bulky”
- Looked different online
    - The context in which these issues have appeared in some reviews, within the dataset published by Brooks (2018), were as follows:
        - Within Tops Department –
            - “nothing like the picture”, “is nothing like what’s shown on the model”, “item, as pictured, is not accurate”
        - Within Dresses Department –
            - “dress shown online is nothing like the dress I received”, “loved it online, but not in-person”

    - Within Bottoms Department –
        - “looks amazing in photos online, but not in-person”
    - Within Jacket Department –
        - “not how it looked on the model”, “looks awful in-person”
    - Within Trend Department –
        - “couldn’t see the issues form the pictures online”

- Price
    - The context in which these issues have appeared in some reviews, within the dataset published by Brooks (2018), were as follows:
        - Within Tops Department –
            - “not worth the money”, “would have kept it if it were much less money”
        - Within Dresses Department –
            - “not worth the price on sale”, “dress isn’t worth the high price tag”
        - Within Bottoms Department –
            - “quality for the price was not there”, “for the amount of price, they do not last a whole year”
        - Within Intimate Department –
            - “price is ridiculous for the item”
        - Within Jacket Department –
            - “not worth the money”

- Above mentioned are some of the contexts in which the respective issues seemed to have appeared frequently.
- Action plan (based on market research) to improve upon the factors identified:
    - Size/Fit Issue - Which size range to keep can be a challenge. Size can be one of the primary reasons for a customer to choose in-store shopping; and companies selling products online mostly get a return from the user because of the size of fit related issues (Why Your Store Needs a Size Chart (And How to Create One), 2016); the researcher provided the following recommendations:
        - The organization should evaluate their existing online portal and check their current sizing guide/chart
        - State the measurements clearly; also include information about height and weight so that customers can identify their right fit even without the help of a measurement tape
        - Make sure the sizing guide is easily visible and made available on every page, this would help the users locate it instantly
        - Make the sizing guide based on the division name; meaning, this company has 3 divisions namely Intimates, General, General Petite, so, sizing guide of each of those divisions should be made available
        - Collecting data from customers by asking close-ended questions specific to size and fit: Is the product of smaller size, or larger size, or is it just right in size?
    - Quality Issue:

- Companies should first understand their customers' definition of quality. With respect to the analysis done on the subset of reviews, it was observed that customers were pointing to the stitching quality, or quality of the buttons or zipper used, or other aspects of the product itself. However, these cannot be the only quality-related issues that will ever exist; collection and analysis of additional reviews might certainly add several other contexts in which this specific issue can arise.
- Islam et al. (2013), in their research, have published a flowchart showing how an organization could review its existing quality system. To be able to resolve quality-related issues that consumers are facing, the company could follow similar steps in order to examine their existing quality system:
    - When a product is getting manufactured, there should be a thorough inspection at every stage.
    - When the fabric comes into the factory, every batch should be checked prior to usage; if it is found to be not according to the set/expected standards, it should be returned to the fabric store. Similarly, additional materials used such as zipper, buttons should also be check; stitching should be exactly how the product is visible online to the customers. Measurement checks and packaging checks should also be a part of the quality system.
    - This would ensure that the quality of every product which is sold by the company is improved and maintained.

- According to Takeuchi and Quelch (2014), users' perception of the quality of a product can be influenced by previous experience, if a warranty is provided, the return policy provided, and so on. To exactly understand the perception of the quality of the current and potential customers of this specific organization, close-ended questions specific to quality should be asked: Rate the quality of this product in comparison to other companies.
- The company should ask close-ended questions pertaining to both product quality and service quality. The analysis done on the subset of reviews shows only product quality-related issues, it is essential to understand what the users think about the service quality as well.

- Material-related Issue:
  - When customers are purchasing clothing items online, they should be clearly coached on how to understand the fabric or material of a specific product. This would be a great feature that will allow the customers to understand the quality of the fabric.
  - Customers can be shown how to understand the material used in a product by making use of the following ways (How do you judge the quality of fabrics when you shop online?, 2016):
    - Features such as the ability to be able to zoom on the product photo and rotate it to see it from all angles, or a detailed video of the product can help the customer look at the patterns, shape, material thickness.

        - Information about the required maintenance of the fabric should be clearly stated for each product; this will help customers to clearly understand if only dry-cleaning is required, or only handwash, and so on. Unless this is specified by the company and followed by the user, the material quality may get degraded.
        - Other factors such as the tensile weight of the fabric used, or if extra buttons are being provided with the item or details about the lining and stitching of the product should also be provided with each item.
- Issue of difference between the look of the product as seen on portal versus the real product that the customer receives:
    - The subset of the data that was analyzed showed that many customers had mentioned their concern/issue pertaining to how different a product looked on the online portal on the model, from the product they received.
    - According to an article, Solve the 7 Biggest Problems of Online Shoppers (2019), the way an organization could go about resolving these issues would be:
        - The company should include images of different models of various shapes and sizes wearing every product
        - Include the specifications pertaining to size, height, weight for each product – this would help the customers know exactly what they should expect while purchasing an item, and how a product will look on them.
- Issue related to the price of the issue:

- In comparison to other issues which were identified based on the analysis performed on the reviews, very few reviews mentioned the price of the product as a major issue.
- Most of the reviewers who did mention price of the clothing item as an issue, it was in the context of the quality of the product; meaning, the item was not worth the price.
- This means, if the other issues are tackled first, it would lead to automatically the resolving of this issue too.
- Current packaging of products can be investigated, to see if there is scope to reduce costs pertaining to packaging; similarly, shipping costs of a product should also be examined (Sheehan, 2017)
- For the costs that the organization incurs, tasks such as entry of data collected from the portal to the right format, basic analysis of the data, and so on should be automated. According to Sheehan (2017), automation of such tasks would help the organization reduce incurred costs by saving time and effort spent; this would in turn help to focus on monitorization of data which would be very helpful for the organization.

4.5 Control Phase

This is the last phase of the DMAIC methodology which, according to Laux et al. (2017), should include: implementation of the recommendations from the Improve phase, so that constant improvement becomes a part of the organization; also, how the already made improvements can be sustained by the organization should be discussed in this phase. For the purpose of this research study, the suggestions/recommendations from the researcher's end have not been implemented by

the e-commerce organization yet. So, while this phase should show how the improvement is being controlled or sustained, the researcher has suggested the means by which sustainability could take place for this project.

- Working on a larger dataset is absolutely required for the e-commerce organization, because only then would the main root issue be identified appropriately after thorough analysis. So while sustainability and constant improvements are the ultimate goals, this would be possible if the organizations collect more data from their side; here, the data collection should include not just additional reviews, but also additional information about their customer base, as well as, taking feedback from customers upon implementation of any change made by the organization. This would help with ensuring that continuous improvement is taking place.
- It is not just a greater number of reviews that are required for deep analysis, but the organization should gather additional information about and from the demographic as well. This means information such as gender, size, or weight about their customer base could help with the following analysis:
    - Considering gender of the individual buying the product can help the organization understand preferences of products by grouping the general purchasing behavior of the respective gender; and this would further help with the branding of different products to attract a specific customer base.
    - While few reviewers have mentioned their size &/ weight in their reviews, the total number of such reviews in the dataset are very few in comparison to the whole population. Getting this information would help with understanding the product development of the company by answering the questions such as: “Are there plus

sizes of clothes available in all of the departments?", "Compare the performance of the different sizes of products on the basis of their sales and reviews; where should the organization focus their attention on?", "What are the major issues faced by customers purchasing different sizes of clothes?", and so on.

- Surveys and questionnaires containing specific close-ended questions pertaining to the issues found should be sent out to current and potential customers before and after implementation of the improvement plan. These surveys could include questions such as: How would you rate the fit of this product in comparison to the other company's?; Would you rate the size of the product as small, large, or right the way it is?; and so on. This would help gather information/feedback pertaining to product and service directly from the customer base and would thereby help build relationships between the customer by letting them know that their opinion has been heard and worked upon.
- Data collected from customer feedback should be monitored daily, so data monitoring should be performed by analysts to see if the changes implemented in the improve phase and gaining the results needed, and if the organization is sustaining these changes. The process which is improved should be monitored since this would lead to continuous improvement (Thomson, 1995), which is the aim of the control phase.

This is how the researcher has shown that the framework proposed by Laux et al. (2017) can be made use of for enhancing customer satisfaction in online shopping portals – which was the first research question (RQ1) of this research study.

CHAPTER 5. CONCLUSION & DISCUSSION

The purpose of this study was to test the framework proposed by Laux et al. (2017), which combines Big Data and Six Sigma, on the e-commerce online shopping portal dataset to see how customer satisfaction could be enhanced. This chapter contains the conclusion of the results achieved from the study, along with the discussion and future research work.

5.1 Conclusions

The research study conducted was for primarily applying and testing the framework proposed by Laux et al. (2017), using both Six Sigma and Big Data techniques, to an e-commerce online clothing store to improve their overall online customer experience. The deliverables for each of the DMAIC framework phase included: Defining the problem statement, goal, motivation, team structure in the Define phase with the help of a project charter; Stating the basic performance measures, and a detailed data-collection plan in the Measure phase; In the Analysis phase - performing exploratory data analysis to explore the dataset fields and relationships, using sentiment analysis to understand polarity of the reviews assigning a sentiment score to every review, generating a linear regression model to examine which factor has the most impact on the opinions of the customer; Recommending improvement plan/solutions in the Improve phase; Suggestions on how improvements could be sustained in the Control phase.

Based on the analysis conducted, the top issues within the customer reviews were found to be related to be size/fit, quality, material, look of the product online versus in-reality, and price. Further, the linear regression model proved that the field Rating had the most impact on the sentiments of the customer; meaning, it aligns with how a customer is feeling about a specific product and could be used for customer decision making. It could therefore be analyzed thoroughly

for enhancing customer experience over time. The results attained thus align with the results of the research study by Poston and Speier (2005), where the authors stated that ratings help with reviewing content and evaluating customer decision making. Also, by the linear model generated in the Analysis phase, it was found that Rating was statistically significant, and had the most impact on sentiment score; meaning, the ratings provided do represent the reviews in the dataset. Therefore, this study on the dataset of the e-commerce organization agrees with the statement provided by Chevalier and Mayzlin (2003) in their research, where the authors have stated that ratings can help "measure the valence of comments without analyzing the comments themselves" (p. 7); in this case, it would mean that since ratings are representative of customers' sentiments, only ratings provided could be analyzed without analyzing the reviews themselves. Furthermore, like Laux et al. (2017) took an example of implementing this framework to higher educational institution, this research study has successfully tested the application of this framework to an e-commerce organization's data of an online clothing store, showing how issues faced by customers could be identified, and which strategies could be implemented to enhance customer satisfaction.

By application of the big data analytics framework (proposed by Laux et al. (2017)), the underlying questions with respect to this research were answered; moreover, the researcher could gain in-depth understanding of the limitations within the data itself. Thus, the framework proposed by Laux et al. (2017) was very comprehensive; for this research study, the framework provided a structured approach for enhancing customer experience with the intention of improving customer satisfaction, by making use of the Six Sigma methodology and Big Data techniques used within every phase of the framework.

5.2 Discussion and Future Work

This research study was conducted to test the framework proposed by Laux et al. (2017) to the E-commerce domain – more specifically, it was applied to a customer reviews dataset of an online clothing store. This framework was based on the DMAIC methodology of Six Sigma, and Laux et al. (2017) had proposed the incorporation of big data techniques in the phases of this methodology.

Linderman et al. (2005) stated the effectiveness of setting goals while working on a project; because the framework proposed by Laux et al. (2017) was adopted by the researcher, the goals of the project were mentioned in the first phase (Define phase) of the framework itself. The outcomes of the Define phase helped clearly understand the project expectation and process flow, with the help of a project charter. The measure phase was then used for determining a data collection plan, and stating the basic performance measures; followed by the analysis phase, wherein, all of the big data techniques were performed, and the two research questions for his research were answered; followed by the Improve phase, which included recommendations by the researcher based on the issues identified; lastly, within the control phase, only recommendations of steps to maintain sustainability were mentioned.

This framework gave a very structured and systematic approach to solve an industry-related problem because of its DMAIC approach - which gave a clear roadmap to be followed for solving a specific issue and/ improving a specific process. The process of analyzing customer reviews for enhancing online customer experience was considered, by the researcher, for this study. While there have been multiple studies performed within the e-commerce domain, implementation of framework proposed by Laux et al. (2017) was yet to be tested. This proved to be an opportunity for the researcher to see how the combination of techniques within the Six Sigma and Big Data domains could assist in enhancing the customer experience/satisfaction, identified as the KPI for

this research study, which e-commerce organizations strive for. The advantages of implementing this framework was that: the researcher was successful in having a clear picture of the problem; various fields of the dataset were explored; sentiment analysis helped deep dive into the opinion of the customers; linear regression helped understand which field to focus on to get to the core issues faced by the customer; finally, topic modeling helped get down to the issues faced – which the researcher confirmed by also performing manual categorization of a subset of reviews based on the issues identified.

Within the Analysis phase of the framework used, sentiment analysis was performed by the researcher to give a sentiment score to every review in the dataset. For this purpose, AFINN lexicon was used with the tidytext package; this means, the pre-existing dictionary consisting of words and respective positive or negative scores were referred to while generating a sentiment score (mean afinn score) for every review. At the end of the analysis, there were a few outliers found; meaning, there were reviews which had a very high sentiment score, but a very low rating given to it, and, there were some reviews with low sentiment score, but a very high rating given to it. For example, one of the reviews in the dataset was: “Waist is tight, and thighs and legs are not. -not flattering- ankle opening is too small, and they have a funny smell”; this review had a very high sentiment score, but a rating of only 1. If the researcher was to manually look at the review, it would have been considered a negative review, but it was given a high sentiment score because of the afinn dictionary that was used – this is one of the limitations of the algorithm used, that was observed because of the outliers existing. As part of the future work, the researcher will develop their custom dictionary which would be specific to the e-commerce domain and the organization in order to check if there is a difference in the results acquired by performing sentiment analysis. Similarly, with regards to the topic modeling that was performed, by using the LDA algorithm,

was unsupervised since the topics were not pre-decided by the organization. As part of the future work, the researcher will fix up on specific topics and then perform supervised topic modeling to see if there is a huge difference in the topics identified; also further, the recommendations provided in the improvement phase will be implemented. For this study, the researcher focused on negative sentiment reviews for generating topics and understanding the issues faced by the customer; in the future, the researcher will also perform topic modeling on the reviews with positive sentiments to examine the topics with the intention of improving upon factors which people have liked so far. The results acquired, based on the analysis performed in this study, cannot be generalized for the whole population since the dataset used consisted of limited number of reviews. To further improve upon the factors identified, as part of the future work, the researcher will have to collect additional customer reviews from the online portal of the organization and will have to repeat the steps performed within the Analysis phase.

The DMAIC approach followed by Laux et al. (2017) certainly worked to give a higher-level understanding and guided on how one could proceed further by working on the deliverables identified. Just that it had to be catered for and tweaked according to what was required for this specific study; meaning, the researcher had mentioned that implementation of the recommendations was not in the scope of the project, so the researcher did not include the deliverables mentioned within the Control phase. Moreover, in the study considered, once more data had been collected, the next step would be to perform the same analysis on the larger dataset available and to see if the issues of the customers which were same and the ones identified early. In that sense, this step could then be called Verify. Whether to still call the process that would be followed using the framework as DMAIC, or to call it DMADV (define, measure, analyze, design, verify) methodology should be studied further (Cronemyr, 2007). Potential further research in this

area could show how the framework proposed by Laux et al. (2017) could combine the DMAIC and DMADV methodologies; the results acquired by testing this new framework could then be compared to the framework proposed by Laux et al. (2017).

Another point to note is that the DMAIC approach requires one to know the problem in-depth before starting the analysis, because the define phase has a project charter as the deliverable, and the project charter contains the problem statement, goal statement, opportunity statement, scope, list of stakeholders, and timeline – this needs to be signed off by the stakeholders involved before moving onto the next phase. The problem/process considered in this study did not have time as a major factor attached to it, however, when considering domains such as healthcare industry, many times time plays a rather important role; and in such projects, there is limited time to quickly solve an issue by analyzing the data. In such time-sensitive projects, how could this framework be adopted should be studied.

Six Sigma tends to be more statistically inclined, while big data projects are more analytics-driven; the framework proposed by Laux et al. (2017) provides a roadmap to look at a process or problem by combining both the domains. Projects in organizations are analytics-driven, and therefore tend to use more technical big data tools and software available. It would be interesting to see development of a tool which could cater to both Six Sigma projects and big data projects, since this would make the framework and its application streamlined and easily adaptable.

Lavalle et al. (2011) had stated the necessity of using the information produced by the organization to its fullest for gaining the insight required for competing in the market. So, considering this point, in this research study, the researcher used the framework by Laux et al. (2017) for achieving the quantitative indicators related to the KPI identified by the organization – which was customer satisfaction. Many researchers, such as Akter and Wamba (2016), had

investigated and recognized the need to perform adequate analysis of data produced in e-commerce firms. Bilgihan et al. (2016) had stated that there were very few studies conducted to investigate the online shopping experience of consumers. Taking these points as the motivation behind this research study, the researcher tested the framework proposed by Laux et al. (2017) specifically for enhancing the online customer experience and for ultimately increasing customer satisfaction in the e-commerce clothing/apparel organization; this provided a segue into solving a big data project using a Six Sigma methodology, thereby leading to continuous improvement – which is a common goal of all organizations these days.

APPENDIX A: IRB EXEMPTION

This study qualified as exempt from the IRB review – the proof of which is shown below.

Nimita Shyamsunder Atal

From: irb@purdue.edu
Sent: Friday, November 22, 2019 1:10 PM
To: Nimita Shyamsunder Atal; Laux, Chad; Springer, John A
Subject: IRB-2019-658 - Initial: Initial Submission - Exempt

Follow Up Flag: Follow up
Flag Status: Flagged

PURDUE
UNIVERSITY

This Memo is Generated From the Purdue University Human Research Protection Program System, Cayuse.
Date: November 22, 2019
PI: JOHN SPRINGER
Department: PWL COMPUTER INFO & TECH
Re: Initial - IRB-2019-658
Applications of Big Data Analytics Framework for Enhancing Customer Experience on Shopping Portals

The Purdue University Human Research Protection Program (HRPP) has determined that the research project identified above qualifies as exempt from IRB review, under federal human subjects research regulations 45 CFR 46.104. The Category for this Exemption is listed below . Protocols exempted by the Purdue HRPP do not require regular renewal. However, The administrative check-in date is --. The IRB must be notified when this study is closed. If a study closure request has not been initiated by this date, the HRPP will request study status update for the record.

Specific notes related to your study are found below.
Decision: Exempt
Category: Category 4. Secondary research for which consent is not required: Secondary research uses of identifiable private information or identifiable biospecimens, if at least one of the following criteria is met:

(i) The identifiable private information or identifiable biospecimens are publicly available;

(ii) Information, which may include information about biospecimens, is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained directly or through identifiers linked to the subjects, the investigator does not contact the subjects, and the investigator will not re-identify subjects;

(iii) The research involves only information collection and analysis involving the investigator's use of identifiable health information when that use is regulated under 45 CFR parts 160 and 164, subparts A and E, for the purposes of "health care operations" or "research" as those terms are defined at 45 CFR 164.501 or for "public health activities and purposes" as described under 45 CFR 164.512(b); or

(iv) The research is conducted by, or on behalf of, a Federal department or agency using government-generated or government-collected information obtained for nonresearch activities, if the research generates identifiable private information that is or will be maintained on information technology that is subject to and in compliance with section 208(b) of the E-Government Act of 2002, 44 U.S.C. 3501 note, if all of the identifiable private information collected, used, or generated as part of the activity will be maintained in systems of records subject to the Privacy Act of 1974, 5 U.S.C. 552a, and, if applicable, the information used in the research was collected subject to the Paperwork Reduction Act of 1995, 44 U.S.C. 3501 et seq.

REFERENCES

Agarap, A.F., & Grafilon, P. (2018). Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network (RNN). *ArXiv, abs/1805.03687*.

Akter, S., & Wamba, S. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets*, 26(2), 173-194.

Antony, J. (2012). A SWOT analysis on Six Sigma: some perspectives from leading academics and practitioners. *International Journal of Productivity and Performance Management*, *61*(6), 691–698. doi: 10.1108/17410401211249229

Bagheri, A., Saraee, M., & Jong, F. D. (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems, 52*, 201–213. doi: 10.1016/j.knosys.2013.08.011

Behara, R. S., Fontenot, G. F., & Gresham, A. (1995). Customer satisfaction measurement and analysis using Six Sigma. *The International Journal of Quality & Reliability Management, 12*(3), 9. doi: http://dx.doi.org/10.1108/02656719510084745

Bilgihan, A., Kandampully, J., & Zhang, T. (2016). Towards a unified customer experience in online shopping environments. *International Journal of Quality and Service Sciences*, 8(1), 102-119.

Blackburn, M., Alexander, J., Legan, J. D., & Klabjan, D. (2017). Big Data and the Future of R&D Management. *Research-Technology Management*, 60(5), 43–51. doi: 10.1080/08956308.2017.1348135

Borkar, V., Carey, M. J., & Li, C. (2012). Inside "Big Data management". *Proceedings of the 15th International Conference on Extending Database Technology - EDBT 12*. doi: 10.1145/2247596.2247598

Brandt, D. R., & Reffett, K. L. (1989). Focusing on customer problems to improve service quality. *Journal of Services Marketing*, 3(4), 5–14. doi: 10.1108/eum0000000002495

Brooks, N. (2018, February 3). Women's E-Commerce Clothing Reviews. Retrieved from https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews.

Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment Classification of Consumer-Generated Online Reviews Using Topic Modeling. *Journal of Hospitality Marketing & Management*, *26*(7), 675–693. doi: 10.1080/19368623.2017.1310075

Chevalier, J. A., & Mayzlin, D. (2003) The effect of word of mouth on sales: Online book reviews. No. w10148. *National Bureau of Economic Research*.

Complete Thematic Analysis with NVivo: NVivo. (n.d.). Retrieved from https://www.qsrinternational.com/nvivo/enabling-research/thematic-analysis.

Cronemyr, P. (2007). DMAIC and DMADV differences, similarities and synergies. *International Journal of Six Sigma and Competitive Advantage*, *3*(3), 193. doi: 10.1504/ijssca.2007.015065

Cudney, E. A. (2012). *Design for Six Sigma in product and service in development: applications and case studies*. Boca Raton, FL: CRC.

Denny, M. J., & Spirling, A. (2018). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, And What to Do About It. *Political Analysis*, *26*(2), 168–189. doi: 10.1017/pan.2017.44

Dogan, O., & Gurcan, O. F. (2018). Data Perspective of Lean Six Sigma in Industry 4.0 Era: A Guide to Improve Quality. In: Proceedings of the international conference on industrial engineering and operations management Paris

Durkacova, M., Lavin, J., & Karjust, K. (2012) KPI Optimization for Product Development Process. *23: 23rd International DAAAM Symposium*. Austria.

Fang, H., Zhu, Q., & Zhang, J. (2011). An Empirical Analysis of the Impact of Online Reviews on Product Sales in the Chinese Context. *IEEE International Conference on Advanced Information Networking and Applications*. doi: 10.1109/aina.2011.43

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5). doi: 10.18637/jss.v025.i05

Fellows, I. (2018). wordcloud: Word Clouds. R package version 2.6. https://CRAN.R-project.org/package=wordcloud

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, *8*(15). doi: 10.18637/jss.v008.i15

Gaffar Khan, A. (2016). Electronic Commerce: A Study on Benefits and Challenges in an Emerging Economy. *Global Journal Of Management And Business Research,* Retrieved from https://journalofbusiness.org/index.php/GJMBR/article/view/1918

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144. doi: 10.1016/j.ijinfomgt.2014.10.007

Gangolly, J., & Wu, Y.-F. (2002). On the Automatic Classification of Accounting concepts: Preliminary Results of the Statistical Analysis of Term-Document Frequencies.

Gentile, C., Spiller, N., & Noci, C. (2007). How to sustain the customer experience: an overview of experience components that co-create value with the customer. *European Management Journal*, 25(5), 395–410. https://doi.org/10.1016/j.emj.2007.08.005

Grün, B., & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 1-30. doi:10.18637/jss.v040.i13 (URL: https://doi.org/10.18637/jss.v040.i13).

Gupta, S., Modgil, S., & Gunasekaran, A. (2019). Big data in Lean Six Sigma: a review and further research directions. *International Journal of Production Research*, 1–23. doi: 10.1080/00207543.2019.1598599

Hayes, D. S. (2000). 1999 International Student Paper Award Winner: Evaluation and Application of a Project Charter Template to Improve the Project Planning Process. *Project Management Journal*, *31*(1), 14–23. doi: 10.1177/875697280003100104

Hernández, B., Jiménez, J., & Martín, M. (2010). Customer behavior in electronic commerce: The moderating effect of e-purchasing experience. Journal of Business Research, 63(9), 964-971.

Hipson, W. E. (2019). Using sentiment analysis to detect affect in children's and adolescents' poetry. *International Journal of Behavioral Development*, 43(4), 375–382. doi: 10.1177/0165025419830248

Hong, L., & Davison, B. D. (2010, July 1). Empirical study of topic modeling in Twitter. Retrieved from https://dl.acm.org/doi/10.1145/1964858.1964870

How do you judge the quality of fabrics when you shop online? (2016, January 29). Retrieved from https://40plusstyle.com/how-do-you-judge-the-quality-of-fabrics-when-you-shop-online/

Islam, M.M., Khan, A.M., & Khan, M.M.R. (2013). Minimization of Reworks in Quality and Productivity Improvement in the Apparel Industry. *International Journal of Engineering and Applied Sciences*, vol. 1, 2013, pp. 147-164.

Jebb, A. T., Parrington, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of Inductive research. *Human Resource Management Review*, *27*(2), 265–276. doi: 10.1016/j.hrmr.2016.08.003

Kuang, X., Chae, H.S., Hughes, B., & Natriello, G. (2017). A Topic Model and Social Network Analysis of a School Blogging Platform. *EDM*.

Lakshmi, S., Gowri, S., Kherajani, M., Jeshnani, H., & Khedkar, S. (2016). A Proposed Framework for Measuring Customer Satisfaction and Product Recommendation for Ecommerce. International Journal of Computer Applications, 138(3), 30–35. doi: 10.5120/ijca2016908757

Laux, C., Li, N., Seliger, C., & Springer, J. (2017). Impacting Big Data analytics in higher education through Six Sigma techniques. *International Journal of Productivity and Performance Management*,66(5), 662–679. doi: 10.1108/ijppm-09-2016-0194

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, 52, 21–31.

Linderman, K., Schroeder, R. G., & Choo, A. S. (2005). Six Sigma: The role of goals in improvement teams. *Journal of Operations Management*, *24*(6), 779–790. doi: 10.1016/j.jom.2005.08.005

Littlejohn, J. (2018, November 30). Retrieved from https://rstudio-pubs-static.s3.amazonaws.com/448239_50a0e8ceb4434fa99178474595674dce.html#sentiment-analysis-other-lexicons

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: the next frontier for innovation, competition and productivity*. New York: McKinsey & Company.

Mehrjerdi, Y. Z. (2011). Six-Sigma: methodology, tools and its future. *Assembly Automation*, 31(1), 79–88. doi: 10.1108/01445151111104209

Meyer, D., Hornik, K., & Feinerer, I. (2008) Text Mining Infrastructure in R. *Journal of Statistical Software*, 25 (5). pp. 1-54. ISSN 1548-7660

Michel, V. (2019, May 24). Sentiment Analysis of IKEA's Google Maps Reviews in R. Retrieved from https://medium.com/@michel.vanessa/https-medium-com-michel-vanessa-sentiment-analyis-of-ikeas-reviews-5b1c14f00c2f

Min, H., Yun, J., & Geum, Y. (2018). Analyzing Dynamic Change in Customer Requirements: An Approach Using Review-Based Kano Analysis. *Sustainability*, 10(3), 746. doi: 10.3390/su10030746

Mizumoto, A., & Plonsky, L. (2015). R as a Lingua Franca: Advantages of Using R for Quantitative Research in Applied Linguistics. *Applied Linguistics*, *37*(2), 284–291. doi: 10.1093/applin/amv025

Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, *42*(3), 1314–1324. doi: 10.1016/j.eswa.2014.09.024

Nasibov, E., Vahaplar, A., Demir, M., & Okur, B. (2016). A fuzzy logic Approach to predict the best fitted apparel size in online marketing. *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*. doi: 10.1109/icaict.2016.7991773

Nielsen, F.Å. (2011). A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. *#MSM*.

O'Cathain, A., Murphy, E., & Nicholl, J. (2007). Why, and how, mixed methods research is undertaken in health services research in England: a mixed methods study. *BMC Health Serv Res* **7,** 85 doi: 10.1186/1472-6963-7-85

Poston, R., & Speier, C. (2005) Effective Use of Knowledge Management Systems: A Process Model of Content Ratings and Credibility Indicators. *MIS Quarterly* (29:2), 2005, pp. 221-244

R Core Team (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Rajgopal, S., Venkatachalam, M., & Kotha, S. (2000). Does the Quality of Online Customer Experience Create a Sustainable Competitive Advantage for E-commerce Firms? *SSRN Electronic Journal*. doi: 10.2139/ssrn.242774

Resnick, P., & Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of eBay s reputation system. *The Economics of the Internet and E-Commerce Advances in Applied Microeconomics*, 127–157. doi: 10.1016/s0278-0984(02)11030-3

Rose, S., Hair, N., & Clark, M. (2011). Online Customer Experience: A Review of the Business-to-Consumer Online Purchase Context. *International Journal of Management Reviews*, *13*(1), 24–39. doi: 10.1111/j.1468-2370.2010.00280.x

Saltz, J. S. (2015). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. *2015 IEEE International Conference on Big Data (Big Data)*. doi: 10.1109/bigdata.2015.7363988

Sheehan, A. (2017, January 24). Cutting Costs: 8 Ideas to Lower Retail Expenses Without Killing Product Quality. Retrieved from https://www.shopify.in/retail/ideas-to-kill-more-costs-without-killing-product-quality

Shmueli, G. (2016). Analyzing Behavioral Big Data: Methodological, Practical, Ethical, and Moral Issues. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2736189

Silge, J., & Robinson, D. (2017). *Text mining with R: a tidy approach*. OReilly Media.

Silge, J., & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *The Journal of Open Source Software*, *1*(3), 37. doi: 10.21105/joss.00037

Soltoff, B. (2020, February 18). Topic modeling. Retrieved from https://cfss.uchicago.edu/notes/topic-modeling/

Solve the 7 Biggest Problems of Online Shoppers. (2019, July 16). Retrieved from https://zoovu.com/blog/solve-the-5-biggest-problems-of-online-shoppers/

Srivastava, M., & Kaul, D. (2014). Social interaction, convenience and customer satisfaction: The mediating effect of customer experience. Journal of Retailing and Consumer Services, 21(6),1028–1037. doi: 10.1016/j.jretconser.2014.04.007

Takeuchi, H., & Quelch, J. (2014, August 1). Quality Is More Than Making a Good Product. Retrieved from https://hbr.org/1983/07/quality-is-more-than-making-a-good-product

Thomson, V. J. (1995). Process monitoring for continuous improvement. *Re-Engineering the Enterprise*, 294–301. doi: 10.1007/978-0-387-34876-6_28

Veiga, G. M. da S. (2019). The application of sentiment analysis to a psychotherapy session: an exploratory study using four general-purpose lexicons. Retrieved from http://hdl.handle.net/10400.12/7372

Verhoef, P. C., Lemon, K. N., Parasuraman, A., Roggeveen, A., Tsiros, M., & Schlesinger, L. A. (2009). Customer Experience Creation: Determinants, Dynamics and Management Strategies. *Journal of Retailing*, 85(1), 31–41. doi: 10.1016/j.jretai.2008.11.001

Vivek, S. (2018, August 22). Analyzing Customer reviews using text mining to predict their behaviour. Retrieved from https://medium.com/analytics-vidhya/customer-review-analytics-using-text-mining-cd1e17d6ee4e

Vladimir, Z. (1996). Electronic Commerce: Structures and Issues. *International Journal of Electronic Commerce*, 1(1), 3–23. doi: 10.1080/10864415.1996.11518273

Wang, J., Zhao, J., Guo, S., North, C., & Ramakrishnan, N. (2014). ReCloud: semantics-based word cloud visualization of user reviews. *Proceedings of Graphics Interface 2014*, pp. 151-158.

Why Your Store Needs a Size Chart (And How to Create One). (2016, June 29). Retrieved from https://www.shopify.in/retail/why-your-retail-store-needs-a-sizing-guide-and-how-to-create-one

Yu, C. H. (2017). Exploratory Data Analysis. *Oxford Bibliographies Online Datasets*. doi:10.1093/obo/9780199828340-0200

Zhang, R., & Tran, T. T. (2009). Helping E-Commerce Consumers Make Good Purchase Decisions: A User Reviews-Based Approach. *E-Technologies: Innovation in an Open World Lecture Notes in Business Information Processing*, 1–11. doi: 10.1007/978-3-642-01187-0_1

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17:5-6, 375-381, DOI: 10.1080/713827180. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/713827180

Zhang, Z. (2008). Weighing Stars: Aggregating Online Product Reviews for Intelligent E-commerce Applications. *IEEE Intelligent Systems*, *23*(5), 42–49. doi: 10.1109/mis.2008.95