An Intelligent Approach for Data Analysis and Decision Making in Big Data: A Case Study on E-commerce Industry

***Abstract*—A recent informational phenomenon has emerged as one of the considerable innovations in information systems, commonly referred to as "Big Data". The latter is currently trendy, both in academia and industry, and is used to describe a wide range of concepts, from data extraction, storage, and management, to data processing and analysis using well-known schemas, to extract patterns in hidden relationships in order to make better decisions and to derive new knowledge using analytical techniques and solutions. The technology that enables the potential of big data to be exploited is called "Big Data Analytics". Big data analytics is a major challenge that enables researchers, analysts and business users to make better decisions faster. Big Data became an important part of marketing research and marketing strategies. The e-commerce industry is one of the industries that currently benefits most from the potential of big data collection and analysis. This paper therefore aims to demonstrate the use of big data to understand customers and to improve and facilitate the decision making process. In this research, we apply multiple machine learning (ML) models on large dataset present in the e-commerce area by studying several practical cases on online markets.**

Keywords—Big data; data analytics; decision making; big data analytics; big data analysis; machine learning; marketing; e-commerce

I. INTRODUCTION

The world has become an information society that is highly dependent on data. Due to this technological revolution, millions of people generate huge amounts of data every day, every second due to the increased use of devices (smartphone, iot.). An interesting engine related to this topic mentions that data growth is unlimited. What is the company going to do about data overload? How to manage and furthermore process and analyze all data to extract value? It seems that we have the "Big Data" problem [1]. The term "big data" was coined when the era of big data began. The latter is an era described by rapidly expanding volumes of data, far beyond what most people imagined. The large volume of data does not only allow us to classify this period as one of big data[2], because there have always been larger volumes of data than our ability to work effectively with data has ever existed. What distinguishes the current era of the big data era is that businesses, governments and non-profit organizations have experienced a change in behavior. Before the big data era began, companies placed relatively low value on the data they collected that had

no immediate value. When the big data era began [3], this investment in collecting and storing data for its potential future value has changed, and organizations have made a conscious effort to retain each potential data element.

Our challenge is not to get data [4][5], but to get the right data and use computers to increase our knowledge of the field and identify patterns that we didn't see or couldn't find before. In any fast-moving field such as Big Data, there is always room for innovation. It is widely recognized, and widely supported by research and case studies, that organizations that use data to make decisions over time actually make better decisions, leading to a stronger, more reliable business. With the speed at which data is created increasing at such a rapid rate, the companies responded by retaining all the data they could capture and assessing the future potential of this data higher than in the past.

A key aspect of this topic is that the ability to analyze and extract information from data for big data decision making is currently seen as an important competitive weapon. Traditional mode analytics, in terms of "Big Data", consists of acquiring data that may or may not be necessary for analysis. All of this requires a different approach, architecture or infrastructure, if necessary. The adoption of new technologies requires processing, discovery and analysis of these massive data sets that cannot be processed using traditional databases and architectures due to lack of resources in terms of computing and storage capacity. Today, data generation is increasing dramatically. This brings new challenges [6] for data analysis, storage, processing and how to get useful information from large datasets. The challenge of analyzing large datasets is to find patterns within them. Therefore, Big data Analytics [7][8] is an appropriate solution to current industrial problems.

Its objective is to "turn data into information" for better decision making and its capability can be turned into a competitive advantage [9][10]. Although Big Data has been one of the most popular topics for many years, how to conduct Big Data Analytics effectively is a great challenge for every field. To analyze big data, institutions are increasingly using analytical solutions. These solutions include predictive and prescriptive analytical techniques [11][12], often using AI and ML and Deep learning in particular, to understand and recommend actions based on the analysis of large volumes of data from multiple sources, internal or external.

By applying big data, it is possible to simplify data analysis and understand some of the characteristics of large data sets. In terms of decision-making, big data analysis [13] will be the next challenge for innovation, competition and productivity. Many solutions will emerge to meet all the challenges in this context. Now, the big challenge for companies is not just to store data, but to analyze it, and that analysis has implications for smarter, more effective decisions [14][15]. Advanced analytics provides solutions to "complex problems" in areas such as production, marketing, human resources or distribution, among others.

Although it may appear that Big Data is mostly employed in scientific fields, in the business world it must also be very present when important decisions are made. In the marketing area, it has become a fundamental piece to carry out large-scale campaigns. Big data is now influencing important marketing decisions. This refers to the application and study of complex, large data sets that cannot be processed by traditional data processing applications. The application of the right technology improves the quality of decision making and detail processes [16]. The term Big Data Marketing [24] emerged as a solution to the needs that marketing has posed since its inception on aspects such as market and consumer analysis. This concept refers to the processes, tools, techniques, and technology used to process large amounts of data in real time, allowing us to analyze consumer behavior, for example, in order to develop better strategies and reach a larger number of people who are interested in our product in a more effective and personalized way.

Indeed, the activity in social media, mobile and e-commerce makes us live in a world of Big Data. Big data technology helps us with advanced segmentation, being able to detect completely new areas of interest of customers. And with this, gain insights to create personalized offers, at the right touchpoint, targeted to the right audience and in real time. Artificial intelligence and machine learning [21] help to understand these metrics and create meaningful trends that indicate future changes in marketing and sales strategy.

Customer relationship management (CRM) [32] becomes a fundamental issue for business operation as a result of big data, which makes the marketing focus of enterprises change from products to customers. Currently, the majority of the E-commerce applications revolve around customer segmentation models, retail analytics, and the use of sentiment analysis to improve business decision making. By conducting this research, we apply an intelligent approach based on Machine learning (ML) to these applications making well-informed decision based on comprehensive data analysis in big data area.

This paper aims to answer the following questions: how to reduce large-scale problems to a scale that humans can understand and act upon? What role can big data play in digital marketing success? And how big data can help e-commerce industry to better understand customers? Based on concrete studies about online markets.

This paper is organized as follows: Section 1 gives an introduction of the research. In Section 2 discuss the theoretical background of this research by providing a general summary and overview of big data and its application on data analysis and decision making. In Section 3 presents literature review on big data applications about customer in marketing. In Section 4 briefs related work about big data techniques applied to marketing. In Section 5, we show the experimentation and results by giving three different case studies in e-commerce industry to see how customers will act in the future based on their present behavior patterns using machine learning approach. Finally, Section 6 gives the conclusion.

AI. Background

A. Big Data

1) *Definition*: Big Data, the synonym for the intelligent handling of such large and at the same time heterogeneous data volumes, is one of the major challenges of our time. Big Data [1] holds great potential for science and industry. It can bring about a lasting change in the way companies make decisions and in the way they conduct research in a wide range of scientific disciplines. Big Data will create scientific progress and innovations and thus increase the competitiveness of our science and our companies.

At the same time, however, it requires a particularly responsible approach to data and the new intelligent Big Data technologies. Big Data refers to the enormous data growth that many companies are currently experiencing. A relatively established definition of Big Data is based on the 3-V model [2][3]. The 3-V model distinguishes three challenges of data growth: volume, velocity and variety. Volume refers to the growing amount of data. Volumes that are considered "big" are in the range of terabytes and more. Velocity describes the speed at which new data are created, but also the speed at which data can be accessed during analysis Variety describes the scope of diverse types and data sources, which can be more or less structured. Some definitions add other V the characteristics of Big Data. Big data can be defined as: "high-volume information assets or varieties that require cost-efficient and innovative ways of processing information and enable better understanding, manufacturing and automation of decision-making processes‖.

2) Big data challenges: In this section, we mention some important challenges related to big data and discuss in more detail some of the technological issues still open to research.

In the Big Data environment, organizations face the challenge [4][5][6] of integrating raw, unprocessed, real-time updated and extremely complex information. But the key issue is not the ability to collect and store voluminous data. It is not enough to simply enter and store a large amount of data, you need to know how to organize, refine and convert it into relevant information that enables you to take market positions. Raw information has only potential value; it is its analysis and systematization that enables organizations to increase their capacity for innovation. Thus, the processing of large volumes of data requires the following steps:

- Acquisition: The data will come from traditional data sources (EDW, relational databases and transactional data files), and from a large number of unstructured

data sources that can be stored in NoSQL and "in-memory" databases.

- Organization of information: Preparing and processing information in order to obtain the best possible results, and on which advanced analytical techniques can be applied as effectively as possible.
- Analysis: Analyze all information with access to all data using advanced statistical tools such as social and opinion mining, or apply techniques developed with the R programming language, specifically to the design of advanced statistics. From an overall perspective, it would be practical for the analytics provider to offer tools for querying and reporting, data mining, data visualization, predictive modeling, and decision optimization.
- Decision making: Make decisions in real time or as quickly as possible so that they can have a positive effect on the company's activities. This step is inseparably linked to the analysis step, indeed many suppliers offer these tools integrated with the decision tools. The decision must be made in real time based on the results obtained in the analysis, so that the raw data is converted into usable knowledge to be integrated into dashboards, prospective dashboards and visualization tools; and thus, predict the behavior of a product or service to consumers.

These many challenges [7] require long-term research to work with big data. Data analysis is more difficult than just locating, identifying and understanding the data. It is not always possible to extract the collected data, through an extraction process, and then transform it into a structured structure suitable for analysis. In addition, another major data challenge is the incompleteness, scale, timeliness and complexity of the process. For large-scale analyses, the extraction process must be automated, which requires that differences in the structure and semantics of the data be expressed in a computer-readable way so that the data can be analyzed.

Another challenge is the management of Big Data, effectively managing big data is essential to facilitate the extraction of reliable information and to optimize spending. Indeed, good data management is the basis of Big data analytics. Big data management means cleaning data to ensure reliability, aggregating data from different sources and coding data to ensure security and confidentiality. It also means ensuring efficient storage of Big Data and role-based access to multiple distributed endpoints.

However, these challenges must be overcome in order to maximize large data sets, as the amount of information exceeds our operating capabilities. Based on an in-depth reading of several articles [4][5][8] which discuss the authors' opinions and perspectives on data analysis regarding the new opportunities and challenges created by the big data movement.

B. Big Data Analytics

The primary objective of analytics is to enable informed decision-making and the resolution of business problems. Thus, analytics is a knowledge repository composed of statistical and mathematical tools, machine learning algorithms, data management processes such as data extraction, transformation and loading (ETL), and computer technologies such as Hadoop that create value by developing actionable elements from data. Analytics refers to BI & A technologies that are mainly based on data mining and statistical analysis. Most of these techniques are based on the mature commercial technologies of relational DBMS, data warehousing, ETL, OLAP. Data analytics refers to the procedures and activities that are meant to collect and analyze data in order to extract usable information. The results of data analytics can be used to: identify key areas of risk, fraud, error or misuse; improve business processes; verify the effectiveness of processes and influence the decisions of the business.

Advanced analytics [8] are implementations of specific forms of analytics that consist of a set of techniques and related types of tools, typically including predictive analysis, data mining, statistical analysis and complex SQL, although the list covers data visualization, artificial intelligence, natural processing languages and the capabilities of analytical databases such as MapReduce, in-memory database analysis and columnar data stores. Big data analytics does not require data to be clean and standardized. In fact, they make no assumptions about data standardization. Data analytics [9] can analyze many varieties of data to provide insights into models and ideas that are not humanly possible. They use advanced statistics, artificial intelligence techniques, machine learning, deep learning, feedback, and natural language processing (NLP) to exploit data. Data analytics today is influenced by all kinds of devices and social media, such as data from GPS, NFC and RFID chips, barcodes and QR codes, and others within the Internet of Things, or data from social networks (Facebook, Twitter). All of them linked to data transit in all types of businesses such as banking, department stores, media, industries, etc.

Big data analytics [10] is the science and technology of organizing big data, analyzing and discovering knowledge, patterns and information from big data, visualizing and reporting the discovered knowledge to aid decision making. Big data analytics is the application of advanced analytical techniques to operate on large sets of large data.

In reality, what is being done is to unite two fields with their own entity: Big Data as a massive amount of detailed information, and advanced analysis which is actually a set of different types of tools, including those based on predictive analytics, data mining, statistics, artificial intelligence, natural processing languages, etc. Also known as process to create sustainable competitive advantages, by exploiting, on the one hand, the pools of knowledge resulting from the fine analysis of new data sources and, on the other hand, the capacity for anticipation, or even prediction, built up from this analysis. Big Data analytics is the use of analytical techniques applied to large data sets.

Therefore, Big Data analytics is really two things: analytics and Big Data. The first one helps to discover those data that have changed in the business to know how to react; big data must help to turn the challenges produced by the spectacular growth of the Big Data. Analytics is the best way to discover new customer segments, identify the best suppliers, associate products by affinity, understand sales by seasonality, etc.

C. Big Data Analysis

Data analysis and business analysis [13] are long-standing disciplines that have experienced remarkable growth in all areas of knowledge and, in particular, in organizations and businesses, due to the need for tools that analyze data and make effective and efficient decisions. Data analysis has evolved as data volumes have increased. Business intelligence tools have brought together OLAP (online analytical processing), reporting and querying, visualization and, most importantly, data mining technologies with their already established categories of web mining and text mining, and innovative social media data analysis, which has relied on techniques for analyzing feelings and opinions, or opinion mining and sentiment, as it is also known. There are a variety of software tools used in data analysis and methods used. The most used techniques are: queries and reports (querying and reporting), visualization, data mining, predictive data analysis, fuzzy logic, optimization, streaming audio, video or photography, etc.

Data analysis is also considered the science of examining raw data for the purpose of drawing conclusions about the information contained therein. It is used in many industries to enable organizations and businesses to make improvements in the of decisions. This term is widely used in the field of business intelligence, and depending on the software tool manufacturer, it can cover a wide variety of terms: OLAP, CRM, dashboard etc [18]. In the era of large volumes, we can consider five major categories in data analysis:

- Data analytics (analytics) in organizations and companies that analyze traditional data: transactional and operational.
- Web analytics or data traffic analysis on a website.
- Social analytics or social media data analysis (blogs, wikis, social networks, RSS.)
- Mobile analytics on mobile devices in order to analyze the data that send, receive or transit these devices.
- Big Data Analysis or analysis of large volumes of data.

Big Data analysis [14] is done with software tools normally used as part of the discipline of advanced analytics. So the usual tools are:

- Advanced SQL queries.
- Queries and reports (querying and reporting).
- Advanced statistical analysis.
- Data visualization.
- Data mining, text mining, web mining and social mining.
- Analysis and predictive modeling.
- Optimization.
- Awareness raising.
- Dashboards and scorecards.

The technologies associated with Big Data mainly include data warehouses, data marts, NoSQL and "in memory" databases, Hadoop and MapReduce frameworks. Big Data, as recognized by all serious studies on the subject by the consulting firms and tool manufacturers, it is an opportunity more than a problem.

Big data analysis [19] is a very active area of research that has a significant impact on industrial and scientific fields, where it is important to analyze very large and complex data repositories. It has found applications in many sectors through its ability to transform huge amounts of data into information for informed business and operational decisions.

D. Decision Making

Decision-making [15][16] has become a key tool in any organization, moving from a process based on experience and intuition to one that is increasingly established in data analysis. Since the early 1990s, data warehousing systems have provided the ability to capture, debug and explore data for support purposes. This triggered the development of decision support systems. From there efficient and intelligent techniques and algorithms to discover hidden information were developed in large databases, known as data mining and later generalized under the concept of Business Intelligence. The fundamental characteristic of these systems was that they ran on large volumes of internal data.

In the age of information and technology, data analysis is essential to forecast business scenarios. Thus, big data has become the main decision-making tool. The digital age has made a large amount of data available to businesses, making it a major asset for improved decision-making in all areas and sectors. Processing this large volume of data, through advanced analytics of Big Data architectures, helps to improve and automate decision making in a company's day-to-day business processes. Big data represent a fundamental change in the way business decisions are made. Sentiment analysis is probably the most widely used tool for decision making based on social network data. For good decision making [17] it is essential that in addition to being the right one, it must be done in a timely manner and with the minimum cost. With big data tools, the crossing of information from multiple sources is possible, which helps to obtain accurate and varied information for decision making. Data management tools suitable for volumes handled by big data are essential.

Data-centric approaches such as big data and related Business Intelligence and Analytics (BI&A) [18] approaches have recently attracted a lot of attention because of their promise of huge improvements in organizational performance based on new business insights and better decision making. Integrating data-centric approaches into organizational decision-making processes is a challenge, especially with big data, and it is not clear that the expected benefits will be realized. Previous studies [16][17] have identified the lack of research focus on the context of decision-making processes in

data-centric approaches. Using ―big datal to improve decision-making has recently become a highly practice area and active. Business Intelligence, the strategic decision-making tool [20].

E. Applications of Big Data in Machine Learning

Machine learning [22] consists of a set of techniques that enable computer systems to predict, classify, sort, order, make decisions , in general, extract knowledge from data without the need to explicitly define the rules for performing these tasks. It is a subfield of artificial intelligence that aims to teach computers the ability to perform example-based tasks without explicit programming. Big data extracts and processes data to make it available to machine learning algorithms. It can be said that big data is the source of data ingestion for ML and DL [22].

Machine learning takes the data processed by big data and analyzes it to generate business insights or learn to perform certain tasks automatically. Deep learning ingests the most important data from big data to learn about it at much deeper levels and to perform more complex tasks. Our target is to study how big data analytics can assist in segmenting online sales data and how to manage a marketing decision making with sentiment analysis on online customer reviews.

F. Big Data in Marketing Area

Marketing and sales are perhaps the areas of greatest application of big data today [23]. Data is used to gain a better understanding of consumers, their habits, and their preferences. Companies are willing to augment traditional data centers with social media, browsing logs, text analytics and sensor data to get a complete picture of their customer. Although Big Data applications to marketing bring innumerable advantages, it is still complicated for many companies to access them [24]. The main challenges stem from the difficulty of choosing the right information from the vast amount available. In addition, there is the challenge of implementing data analytics that knows how to extract valuable and actionable information. Finally, the biggest challenge has to do with people and lies in ensuring that all levels of the company, including top management, assume a culture based on data analysis.

Big Data is the biggest and best tool that marketers can use for their campaigns and strategies. With Big Data, we can analyze in real time large amounts of data and generate personalized profiles of our customers. This helps us to better understand their interactions with our company and purchase intentions. Big data can make more accurate real-time decisions in marketing. Big data allows companies to better focus on the primary needs of customers by developing quality and informative content. It enables us to study how to do a good job of customer segmentation and clustering, analyze customer demands and needs, optimize marketing strategy distribution models by the use of sentiment analysis to understand how customers feel, improve the real estate marketing system [25][26].

BI. Literature Review

In this Digital Era we have at our disposal a large amount of data that defines the behavior of our customers [23]. Understanding the customer is a key element of the marketing strategy. With the fast changes in the market, marketing strategies based on the change in customer behavior relative to the problems caused over time has been a challenge.

In this section a review of literature on customer segmentation, retail analytics, sentiment analysis.

A. Review on Customer Segmentation

Customer Relationship Management (CRM) [31][32] is a competitive strategy for understanding a company's customers. But the question is: On which customers should efforts be focused to build successful and competitive relationships, considering that not all of them have the same importance for the company? To determine this level of importance, the use of customer segmentation techniques is useful and decisive in identifying those customers who are really profitable, and allows focusing more resources on them, maximizing their value, and also, optimally using resources in terms of attracting, retaining or recovering them.

In 1956, Wendell R. Smith revolutionized marketing by introducing his segmentation theory. Segmentation is the practice of classifying customers into different groups, based on their multidimensional information (socio-demographic characteristics, purchasing and usage patterns, preferences, needs and attitudes). Customer segmentation [27] is crucial to creating a marketing strategy, as it allows you to better understand the composition of the audience and then propose a marketing mix that precisely meets the needs of each user belonging to a targeted segment.

B. Review on Retail Analytics

The most of publications on big data analytics are focused on technical algorithms or systems development. The advantages of using big data analytics are not limited to a particular industry. The retail business has entered the Big Data age. Using Big Data, retailers can identify certain issues through customer activity, feedback, comments, and reviews on social media and reduce the problem by stocking products with higher demand thus increasing sales potential. Today’s retailers are up against a perfect storm of challenges. Consumer expectations and behaviors are changing, and a rapidly evolving consumer decision-making journey. Big data is to blame for a massive transformation in the retail sector [28] .

C. Review on Sentiment Analysis

Sentiment analysis (also known as opinion mining) is the process of determining or measuring the tone, attitude, opinion and emotional state of responses, to decide whether a conversation or opinion is positive, negative or neutral. Its use occurs in environments ranging from airlines, insurance, clothing sales and financial institutions, to political decisions. Sentiment analysis [29] is a common research Topic in the Natural Language Processing field (NLP). Natural Language Processing (NLP) is a field within the area of artificial intelligence, computing and linguistics. Its main objective is to facilitate and make effective communication between people and computers through the use of protocols such as natural languages. It transforms text into a language that the machine can understand, Big Data collects large amounts of data to obtain a more accurate analysis by improving the performance of the algorithms, and Artificial Intelligence (AI) uses the

information provided by NLP to determine the categories of feelings and their corresponding polarities.

One of the techniques applied to large amounts of data is sentiment analysis. Its objective is focused on analyzing the vocabulary of a text in order to determine the opinion that a person has about a certain topic based on the ideas expressed in a text. NLP enables researchers to gather and analyze such data in order to determine the basic meaning of such writings. The field of sentiment analysis, which is used in a variety of fields, is highly reliant on NLP techniques. Opinion mining is a sub-domain of text mining, which consists in analyzing texts in order to extract information related to opinions and feelings (Sentiment Analysis).

IV. Related Work

E-commerce has evolved over the last few decades, and today there are millions of electronic transactions, i.e., purchases that are made over the Internet (online). Our target is to study how big data can aid in segmenting online sales data for better marketing strategies, and how big data analysis can help retailers to make strategic, effective and timely decisions [30].

In this section, we introduce the application of big data in Marketing and its uses for customer segmentation, retail analytics, and sentiment analysis. In this section, we are going to discuss related works and techniques proposed by various researchers that relate to big data analysis in marketing field.

An approach for business customer segmentation that integrates clustering and multi-criteria decision making is developed by [30]. This study extends the traditional RFM model by including five novel segmentation variables for business markets. The results of the application show that the proposed approach for business customer segmentation can effectively be used in practice. By increasing the effectiveness of CRM strategies, the proposed segmentation can help firms gain a sustainable competitive advantage in the market.

Another study [31] aims to determine the best approach to customer segmentation and to deduce associated rules for this based on recency, frequency and monetary (RFM) considerations as well as demographic variables. In this research, the impacts of RFM and demographic attributes have been challenged in order to enrich variables that aid comprehension to customer segmentation. The results show that the weights of RFM attributes have a positive effect on rule association performance. A study [32] examines particular e-commerce enterprises by using cluster analysis and logistic regression analysis to predict and analyze customer segmentation and customer and churn retention.

The aim of this research [33] is to demonstrate how big data analytics can be used in customer segmentation. The authors analyze various data analytics algorithms, especially K-Means and SOM. Although K-Means produced promising clustering results, SOM outperformed in terms of speed, quality of clustering, and visualization. This study discusses how these segmentation analysis approaches can be helpful in studying consumer interests.

A survey research on parallel processing systems [34], gives overview of the existing parallel data processing systems categorized by the data input as stream processing, graph processing, batch processing, and machine learning processing and discusses how the MapReduce architecture works and how efficient it is for data processing. It implements Spark as an easy-to-use cluster computing platform. It includes parallel operation features such as reduce, for each, and collect, which allow users to perform filter, map and reduce operations using functions.

The main focus of this study [35] is to use the Hadoop framework to analyze clickstream data collected from an online retail e- commerce website. In order to handle big data efficiently, the author use a variety of technologies such as Pig, Hive, and Sqoop, which are based on the Map-reduce.

A research [36] uses a big data platform to do sentiment analysis on a huge collection of tweets by examining user social data on a certain topic.

The author in [37] proposed implementation of the Spark framework for analyzing Twitter data. Spark offers and provides efficient solutions to analyze large amounts of data. Spark was able to process the tasks within short time which shows the high efficiency of the framework.

In this research [38], Apache Hive and Apache Pig are used to analyze the performance of various ECG Big Data datasets. Different ECG Big Data factors were examined, and the findings revealed that Apache Pig was more efficient and systematic than Apache Hive in terms of delivering rapid results in less time.

Since 2002, a large number of articles citing Sentiment Analysis have been published, focusing on the classification of comments and their polarity (positive or negative). Many researches have focused on opinion texts about products or political discussions. A research [39] propose a preference based sentiment analysis. It uses Hadoop technology to analyze various products. This paper shows the importance of analysis in a huge amount of data. This study proposed sentiment analysis of tweets, but accuracy and data size are very low.

A research [40] propose sentiment analysis of Twitter data using basic machine learning algorithms. To classify the text data. The author used basic machine learning algorithms such as KNN and SVM.

The aim of this research [41] is to improve marketing intelligence by developing a better understanding of consumers online generated contents in terms of positive and negative feedback. It focuses on the collection of tweets referring to three fast fashion retailers of various sizes operating in the UK market, and then analyzed and evaluated through a sentiment analysis based on machine learning. It provides an efficient and systematic approach to accessing the rich data set on customer experiences based on the vast amount of information that customers produce and share online, as well as investigating this massive amount of data to achieve perspectives that can affect retailers marketing intelligence.

The importance of log files in the E-commerce word is discussed in this study [42]. It proposes a predictive

prefetching system based web log preprocessing using Hadoop MapReduce, which deliver accurate results in the shortest time possible for e-commerce business operations.

V. Experimentation

In this section, we have treated several case studies based on e-commerce industry to demonstrate the role of big data as a major marketing tool in understanding customers and improving the decision making process. Big data is a powerful tool to analyze and obtain very valuable information in any field. Applied to marketing, it allows us to study large volumes of data to extract information about customers, their interests or their consumption habits and thus be able to design more effective strategies. In this way, we are moving to marketing intelligence [43], which uses big data to understand customers, their environment, and make more strategic decisions. Big Data Analysis offers the analysis of large volumes of data to detect relationships between them that can provide useful information to companies, facilitating decision making in all processes and areas of the organization. The potential of big data can be fully exploited by corporate marketing departments to make decisions based on customer data. Machine learning presents a huge growth opportunity for online retailers. With machine learning, smart e-commerce companies can boost sales, reduce waste, and increase overall efficiency while actively engaging with consumers. It is a great tool for making better decisions.

In this paper, the experimentation will be separated in three subsections as follows.

A. *Case study 1 : Approach K-Means and RFM to Study the Evolution of Customer Segmentation in Big Data Area*

Rooted in recent literature [43][44][45], we focused on the landscape of big data analysis through the lens of a marketing mix framework [43]. The digital transformation enables enterprises to mine big data for marketing intelligence on customers, markets, products. The customer segmentation pillar of the marketing strategy. Segmentation is a way of dividing a problem into simpler parts, which help to prioritize efforts and locate business opportunities. Not all is a way of dividing a problem into simpler parts, which help to prioritize efforts and locate business opportunities. Not all customers are the same or have the same needs. Companies must therefore understand this and adapt their value proposition to each target group. Segmentation [45] is the process of dividing a population into homogeneous groups based on needs, behaviors, characteristics or attitudes and characterizing the resulting groups to discover what distinguishes them from one another.

Clustering [50] [51] is a machine learning method widely used in marketing for customer segmentation. A very useful way to segment and understand what kind of population we are studying is to see how their individuals naturally organize themselves. Today, clustering algorithms are commonly used in the commercial field, such as customer analysis, and this application has achieved a positive impact and good effect.

1) Proposed methodology: This research proposes a case study of building a customer segmentation model using data mining methods, K-means clustering, and RFM analysis, we used the k-means [52] and RFM [53] approach to study the evolution of customer segmentation in the era of big data [45][52]. The proposed methodology is shown in Fig. 1.

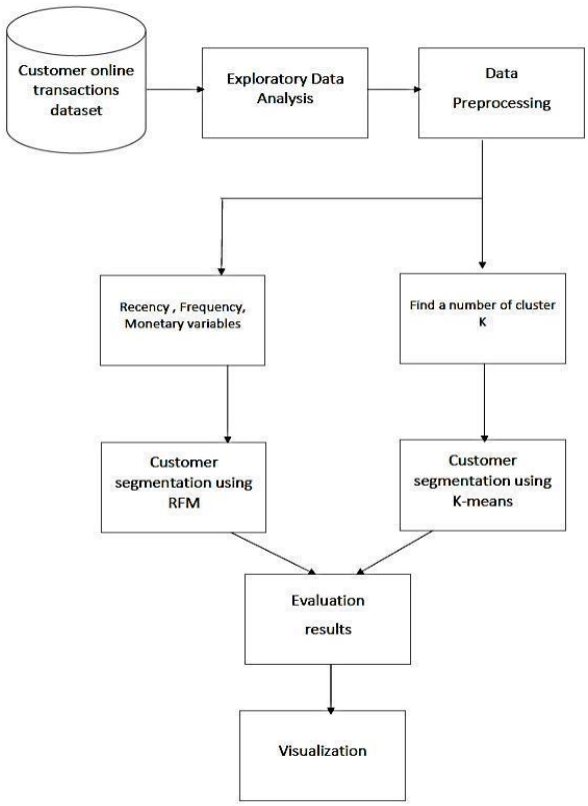


Fig. 1. Proposed Methodology for Customer Segmentation.

2) Dataset: The dataset used in this case study was collected from UCI Machine Learning Repository. This dataset is very hierarchical, with 541909 transactions and 8 attributes that describe these transactions, it contains all transactions recorded over an eight-month period (01/12/2010-09/12/2011) for a UK-based online retail company.

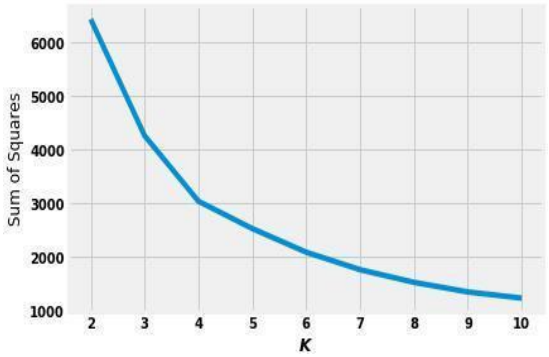
3) Data Pre-processing: In this stage, cleansing of the dataset was performed using Python programming language in Jupyter. Pre-processing of this data collection includes eliminating NAs, validating numerical values, removing invalid data points, and normalizing data.

4) *Method*: To achieve an efficient customer segmentation we, this study starts with K-means clustering [50] by the following steps to find customer groups with similar behaviors:

- Choose the number of clusters k.
- Set the k cluster centroids to their initial values.
- Assign the n data points to the clusters that are closest to them.
- Using the data points in each cluster, update the centroid.

- Repeat steps 3 and 4 until the variations in centroids' positions are zero.

a) *K-Means Algorithm*: K-means [52] algorithm requires knowledge of the number of groups in advance and therefore two internal evaluation methods were applied to determine this parameter: elbow method and silhouette method. In Fig. 2, the result of the elbow method is shown, the X-axis represents the cluster number and the Y-axis represents the sum of squares for the clusters (within-cluster sum of squares (WCSS)) according to. The appropriate cluster solution is defined at the moment when a dramatic reduction of the sum of squares in the cluster occurs. Based on Fig. 2, the extreme decrease of the sum of squares is pointed in the k=4 This produces a "bend" in the frame and in this case this bend can be seen in the number of four clusters. To validate the interpretation of this graph, the number of groups was also obtained using the silhouette index, with the highest index also being obtained in four groups.



these clusters. As shown in Fig. 4 customer segment in Red has low total sales and a low number of orders, which means that they are low value-added customers. On the other hand, Yellow customers have a high total sales and a high number of orders, indicating that they are the customers with the highest value. For Fig. 3, we could look at customers in the red group and try to find ways to increase their orders with e-mail reminders or targeted SMS notifications based on certain other identifying factors. The Fig. 5 confirms the previous two charts by identifying the Yellow cluster as the customers with the highest value, the red as the customers with the lowest value, and the blue and green as the customers with the highest opportunity. Customer segmentation can be carried out by taking into a multitude of different variables, such as demographic, geographical, psychological (preferences, etc.) or purchasing behavior data, which are increasingly important. Based on RFM analysis, Fig. 6 helps us decide which customer groups to target and how we can communicate with them.

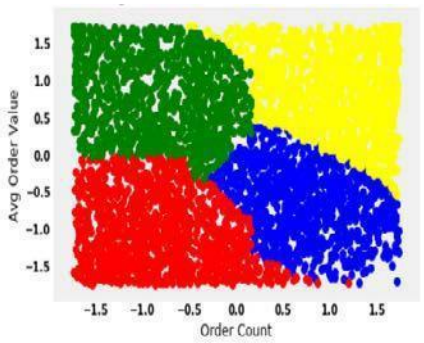


Fig. 2. Elbow Method for k-Means Algorithm.

b) RFM: To build a successful customer relationship management, companies must start with the identification of the true value of their customers as this provides the foundation for implementing more targeted and personalized marketing strategies. RFM (Recency- Frequency – Monetary) analysis is a data driven customer behavior segmentation technique.

RFM [53] based on three metrics that determine their purchase habits: Recency (the date since the last purchase), Frequency (how much the customer makes purchases), and Monetary value (the total value of all a customer's purchases). Customers purchase dates are typically sorted by RFM terms, which are determined by the number of appropriate time intervals. The quantile values are given a top score 4 and other 3, 2, 1. The quantile method is used to sort customer data in descending order (high to low), and the RFM scoring method is referred to as the customer quantile method.

RFM analysis enables businesses to properly forecast which consumers are more likely to make future purchases, how much income comes from new (vs repeat) clients, and how to convert infrequent consumers into habitual ones.

5) Results and disscusions: Based on the results of cluster analysis, 4 is the optimal number of clusters for this analysis. For the interpretation of the customer segments provided by

Fig. 3. Avg Order Value vs Order Count Clusters.

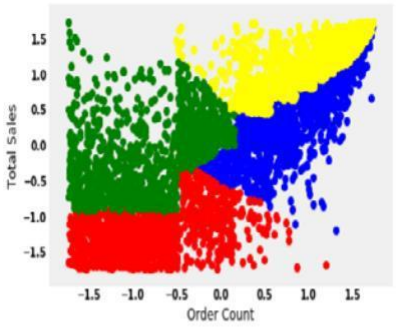


Fig. 4. Total Sales vs Order Count Clusters.

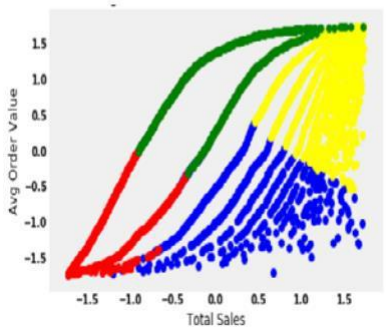


Fig. 5. Avg. Order Value vs Total Sales.

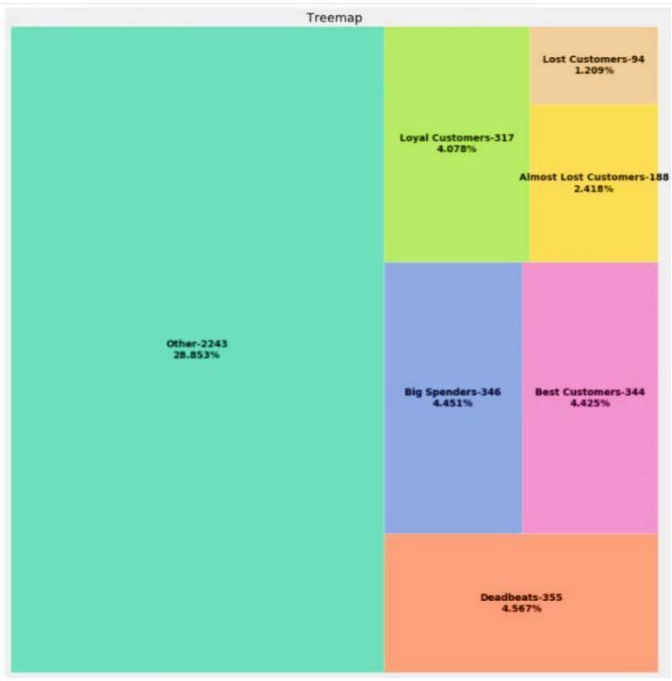


Fig. 6. Summary Report on Customer Segmentation using RFM.

development [54]. Most research on big data in retail has focused on how to gain greater consumer insights to implement marketing activities and how to aid retail business owners in better understanding their customers' buying demands.

In this study, we used a distributed computing approach to analyze and improve retailing business using big data framework. It focused on online retail analytics using the same dataset used in the previous study.

1) *Proposed architecture*: The following proposed architecture Fig. 7 illustrates the analysis of Online retail transactions using Apache pig, Hive, Spark, Apache Zeppelin. For analyzing these large amounts of data a power tool is required using Hadoop and we also need efficient analytical tools which work on the top of Hadoop, Apache hive and Apache pig [54][55]. The first step was to install the necessary tools and the configuration between the set of big data analytics tools in order to build this architecture.

6) Challenges and limitations: There are many challenges identified in this study, this analysis is usually performed in the form of a snapshot analysis where segments are identified at a specific point in time. However, it doesn't take into account the fact that customer segments are highly volatile and that segments change over time. Once the segments change, the entire

analysis must be repeated and the strategies adapted. This is the origin of the grouping of flows as a tool to mitigate this problem. One of the biggest challenges is that customer segmentation is often based on a customer's transaction history.

As this data changes over time [49], it is necessary to update clients that have already been included in the bundle. We suggest the use of Stream clustering[46][47][48] is an extension of traditional clustering that manages a continuous flow of new observations. It updates the underlying classification over time without the need to recalculate the entire model. While it seems promising to apply flow aggregation for customer segmentation, it comes with several challenges. This customer segmentation is based on customer transactions. If we had data on customer demographics (i.e., gender, age, location, annual salary), we might discover more interesting information.

B. *Case Study 2 : A Distributed Computing Approach to Analyze and Improve Retailing Business using Big Data Framework*

Due to the complexity of big data, there are many obstacles to analyzing the dataset. The time span of the analysis is significant and important because it influences how quickly decisions can be taken in response to a change in the business environment. Analyzing the massive amounts of data produced by such transactions is critical for business growth and

Fig. 7. Proposed Architecture based Big Data Tools.

2) Environement: For the analysis, many preliminary settings needed to be completed before moving to code development. The first step was to install Hadoop and install Apache pig, Apache Hive on top of it. Each tool has its own set of requirements, so how these tools are used is determined by the importance of the data and the needs of the organization or business. Due to the large amount of data that we need to be constantly analyzed, we attempted to exploit a physical server. Since Virtual machines are vulnerable to performance problems as a result of an overflow of virtual servers in a physical servers. However, in a large industrial environment, VMs may not be efficient because scalability of the processing equipment can be a problem. The performance should be considered in our case. In fact, Physical servers are much more powerful and efficient than VMs.

- Hive: It would be impossible to analyze massive databases without taking into account the technology's performance. Hive is a robust data warehouse platform built on Apache Hadoop. For big data analysis, the two together have stable storing and processing capability.

- Pig: Pig allows us to analyze and process the large datasets quickly and easily. It is also known as dataflow language. Ensures the originality of data by decreasing coding lines and replication.
- Spark: Apache Spark is another tool that has a large following. One of the great advantages of Spark is that it is able to store a lot of the processing data in memory and on disk, which can be much faster. However, one of the most interesting qualities is its ability to run on a single local machine, making it far more easier to use.

3) Proposed methodology

There are many steps Fig. 8 for the analysis of Online retail: :

- Step 1: Collect consumer transactions datasets from web resources.
- Step 2: Load dataset using Hadoop command line.
- Step 3: Store the data on HDFS which is very reliable for storing complex or large data size.
- Step 4: The data Ingestion, Cleanup and aggregation using Apache Pig, and SQL on Hadoop using Hive.
- Step 5: Analyze data and manipulate SQL in Hadoop using Hive. Also the analysis and visualization using SparkSQL on Apache Zeppelin.

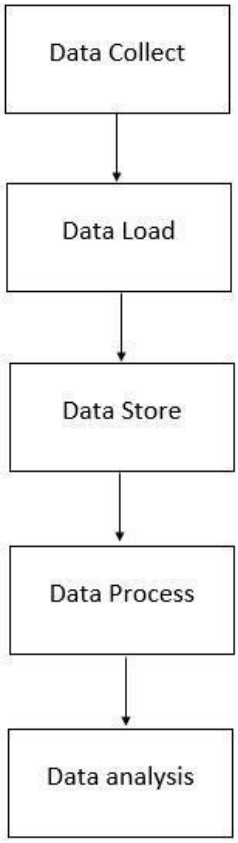


Fig. 8. Proposed Methodology to Analyze Retailing Business.

4) Result and disscusion: The below are some of the problem statements that have been analyzed in this study:

- The sales report by analyzing daily and hourly sales activities is done; this can help in decision making and aid in finding potential new market opportunities.
- To increase sales, an analysis of Basket size distribution is done. It aids in the identification of target audiences, the acquisition, retention, and growth of customers.
- In order to enable online marketers and online retails to build their business strategy around the most valuable customers, an analysis of customer lifetime value is done.

This study experimentally validates the effectiveness of the proposed architecture for retailing business. Big data analytics using pig and Hive [55] highlight major problems faced by consumers and help institutions or companies to solve these problems, provide appropriate consumer satisfaction, improve services, monitor problems and strengthen goodwill in the market place.

Query languages like as Hive and Pig became popular in the analysis of customer data to ensure that customers continue to have a convenient experience and businesses continue to provide excellent services. Two major factors make big data analysis the most meaningful leap in the history of data analysis. The low cost of computer hardware and availability make the analysis of big data the current trend of this era [54]. Large datasets amount to terabytes and require very complex operations and algorithms to be executed, but results are achieved in minutes, sometimes even seconds [56].

C. Case Study 3: A Machine Learning Approach for Statistical Analysis on Customer's Reviews with Sentiment Classification

1) *Motivation and background:* Consumers online posts, interactions, rating and ranking, reviews of products/ restaurants/attractions provide a large amount of data that marketers might access to enhance and improve the decision-making process, by influencing the competitive and marketing intelligence. Sentiment analysis is a contextual study that tries to determine people's feelings, views, outlooks, emotions, and moods regarding entities and their features. Companies are starting listen to social media as a tool to understand their customers in order to further improve their products and/or services. As part of this trend, text analysis has become an active research field in computational linguistics and natural language processing. One of the most popular problems in this area is the classification of texts . Today, understanding how customers feel is important key in marketing strategies. Sentiment analysis focuses on the understanding of emotions and the analysis in text patterns. Sentiment analysis in machine learning is a natural language processing (NLP) problem. NLP [57] is a field of artificial intelligence related to the understanding and processing of language. Both methods are very important for companies as they determine the public

reaction to certain topics, products and/or services on digital platforms.

The aim of this study is to use sentiment analysis to find an accurate classification method for customer reviews based on online women clothing reviews by building a classification model to predict whether the customer will recommend the product or not. We used the power of text mining [58][58] to do an in-depth analysis of customer reviews.

2) Dataset: For this study, we used a real commercial data Women's e-commerce clothing reviews dataset revolving around the reviews written by customers. To maintain the data privacy, company name is replaced with word ―retail‖ and customer names are excluded. It includes 23,486 rows and 10 feature variables in one CSV file. Each row corresponds to one customer review. The choice of the clothing e-commerce company as a population because this market is very interesting and attractive.

This application attempts to understand the correlation of different variables in the reviews and opinions of customers on a women's clothing e-commerce, and to classify each review according to whether or not it recommends the product under review and whether it is a positive, negative or neutral feeling.

3) Proposed methodology

This study was accomplished through a series of steps by following the entire process Fig. 9.

- Step 1: Collect real-time customers' data from Kaggle.
- Step 2:Analyze and plot each of the features in the dataset, this step covers four statistical analysis and visualization to gain insights on the features into the dataset and generate our hypotheses.
- Step 3: This step involves a lot of steps. Cleaning of data by removing outliers, Data Encoding, handling missing values and removing redundant features.
- Step 4: This step comes under data preprocessing. It includes cleaning and preparing text data to reduce the unnecessary characters by eliminating delimiters, to extract stop words, formatting and removing tokenization, Normalization.
- Step 5: Sentiment analysis is the process of analyzing customer sentiment with the use of NLP, text analysis, and statistics. It is accomplished by the use of algorithms to identify words as positive, negative, or neutral. This analysis tells us the polarity score of online reviews. Polarity is a float ranging from -1 to 1. Using the polarity score, we classify each of reviews into these three categories (positive, negative, neutral). Word cloud is used to extract the most used words and classify their polarity.
- Step 6: This step contains our classification models that are built using a detailed analysis of customer review text data as well as other numerical, categorical data.

For our study, we used four models Decision Tree, Logistic Regression, Support Vector Machine and Naive Bayes to build a sentiment classifier.

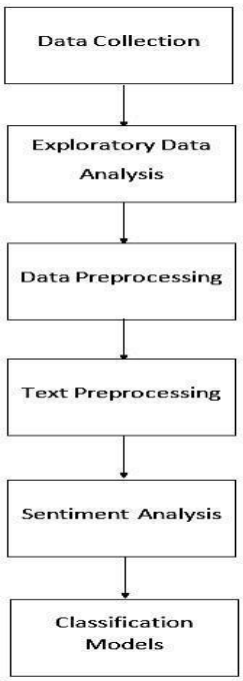


Fig. 9. Proposed Methodology for Sentiment Analysis.

4) Result and disscusion: In this experiment, we used supervised and unsupervised techniques [59] to predict the customer sentiment from reviews. We build our classification models based on a detailed analysis of customer review text data as well as other numerical, binary, and categorical data.

As previously stated, the target variable here is whether or not the customer will recommend the product. We utilized four models: SVM, Logistic Regression, Decision Tree and Naive Bayes. To evaluate the performance of our model accuracy, precision, F1-score, recall scores calculated. (1)

= (2)

(3)

1 (4)

TABLE I. ALGORITHMS AND METRICS

Algorithm	Metrics			
	Accuracy	*Precision*	*Recall*	*F1-score*
SVM	0.95	0.98	0.92	0.95
Logistic Regression	0.95	0.98	0.93	0.95
Decision Tree	0.95	0.98	0.92	0.95
Naive Bayes	0.65	0.86	0.36	0.51

The Table I above showed the experimental results, it was obvious that expect Naive Bayes all other Classifier models perform well in Accuracy and F1-score as well as other metrics. The accuracy, precision and F1-score was the same as SVM, Logistic regression, Decision tree, which was 95%, 98% and 95%, respectively. However, Logistic regression had the highest Recall of four algorithms, which was 93%. In classification, naive Bayes converges faster but has greater error rates than other models. On small datasets, Naive Bayes is preferred, but as the training set size rises, other models are likely to outperform Naive Bayes. Based on the accuracy and F1 score in the chart Fig. 10, we can conclude that, with the exception of Naive Bayes, all classifiers can be utilized to analyze customer reviews.

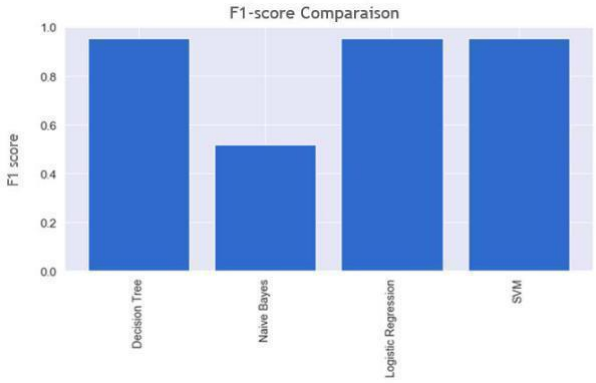


Fig. 10. Performance Comparaison of Models.

VI. Conclusion and Future Work

Big Data, as an emerging and continuing trend, is widely considered, studied and implemented in various sectors. While the level of implementation varies from sector to sector, it is important to understand what it takes to implement, reap the benefits of using Big Data. Big Data in e-commerce acts as the intelligent system that helps in the understanding of data and its transformation into valuable information, allowing us to better control all the processes of the online business. With the help of Big Data, online retailers may get more accurate information about the market, their users, and their customers, making it much easier to make decisions that will help them increase their return on investment. Otherwise, the ability to analyze and extract information from big data is currently seen as an important competitive weapon. Big data analytics is the field in which advanced analytical techniques operate on big data. These techniques are based on algorithms that help us obtain insights, patterns, correlations and associations that could not be understood with traditional small data. It also represents a real advance in the quality and customer service that any e-commerce company can and should offer. In the future, e-commerce will profit from efficient Big Data analysis and processing.

In conducting this research, it became clear that Big Data can improve decision making especially in e-commerce, it is critical for personalizing the strategies implemented, creating dynamic pricing, improving customer support before, during, and after transactions, better managing purchases, and making predictions. Big data has been responsible for some of the most

significant innovations that have revolutionized the e-commerce industry.

An intelligent approach for data analysis and decision making in big data is demonstrated using a marketing cases study based on e-commerce industry by applying several machine learning models. The objective of this paper, as mentioned in the introduction, was to obtain and review knowledge about data science, particularly the field of big data, as well as to explore the existing literature on the fundamental concepts of big data and big data analytics and also demonstrate the role of big data as a major marketing tool in understanding customers and improving the decision making process based on practical case studies on online markets.

This research was done by a various case studies on e-commerce to study the evolution of customer segmentation in big data area using Online retail dataset with k-means and RFM, to analyze and improve retailing business using Big data framework, and to find an accurate classification method for customer reviews based on online women clothing reviews. Research on big data and their potential value in e-commerce industry is still very limited. When it comes to leveraging data, the volume of data, and the speed at which it accumulates the variety of data are the biggest challenges for e-commerce enterprises to overcome. To effectively utilize the benefits of Big Data in E-commerce, advanced research should be conducted in order to overcome application and existing technological challenges. The set of applications carried out has helped us to discover and better understand this topic, in particular the contribution of big data analysis and decision making in big data by looking for challenges and perspectives in e-commerce industry.

In addition to this research, the future work should be focus on including new data integration tools, new reporting tools, and new querying tools which can work efficiently with the challenges of big data.