UNIVERSITAT DE BARCELONA
BUSINESS SCHOOL

MSc
Business Research

MASTER THESIS

Big data and Sentiment Analysis considering reviews from e-commerce platforms to predict consumer behavior

MSc IN BUSINESS RESEARCH

University of Barcelona

Author: Sergi Pons Muñoz de Morales

Directors: Javier Manuel Romaní Fernández & Jaime Gil Lafuente

Barcelona, 16th June 2020

Contents

Abstract

Nowadays and since the last two decades, digital data is generated on a massive scale, this phenomenon is known as Big Data (BD). This phenomenon supposes a change in the way of managing and drawing conclusions from data. Moreover, techniques and methods used in artificial intelligence shape new ways of analysis considering BD. Sentiment Analysis (SA) or Opinion Mining (OM) is a topic widely studied for the last few years due to its potential in extracting value from data. However, it is a topic that has been more explored in the fields of engineering or linguistics and not so much in business and marketing fields. For this reason, the aim of this study is to provide a reachable guide that includes the main BD concepts and technologies to those who do not come from a technical field such as Marketing directors. This essay is articulated in two parts. Firstly, it is described the BD ecosystem and the technologies involved. Secondly, it is conducted a systematic literature review in which articles related with the field of SA are analysed. The contribution of this study is a summarization and a brief description of the main technologies behind BD, as well as the techniques and procedures currently involved in SA.

JEL codes: C38, C45, C53, C55, C63, C67, C88, D12

Introduction

It is not new that that in recent times users have been making more and more intensive use of digital technologies. This usage carries with it an increasing generation of data by users and machines. This huge amount of digital data also known as Big Data (BD) has changed the companies' competitive environment. On the one hand, new affordable tools have appeared democratizing Big data management regardless of company's size (Jelonek, 2017). On the other hand, these tools implicitly include complexity for organizations.

In addition, as the amount of the data generated by users increases, due to the improvement of internet connections worldwide, the challenges to handle this amount of data increase too. For this reason, tools like machine learning (ML) can help organizations to handle and take advantage of data generated by users.

One type of data provided by users with high value, either for other users or for the sellers, is the reviews of purchased items made in the e-commerce marketplaces. Many users in internet rely on reviews of other users to decide whether to make the purchase

or not; this fact supposes a challenge and opportunity for the sellers to improve their products and track consumer behaviour (Statista, 2020).

BD has been a buzz topic for the last few years. But, when did it start to spark interest? According Google trends (2020), the searches of the term Big Data within the last 16 years are as follows:

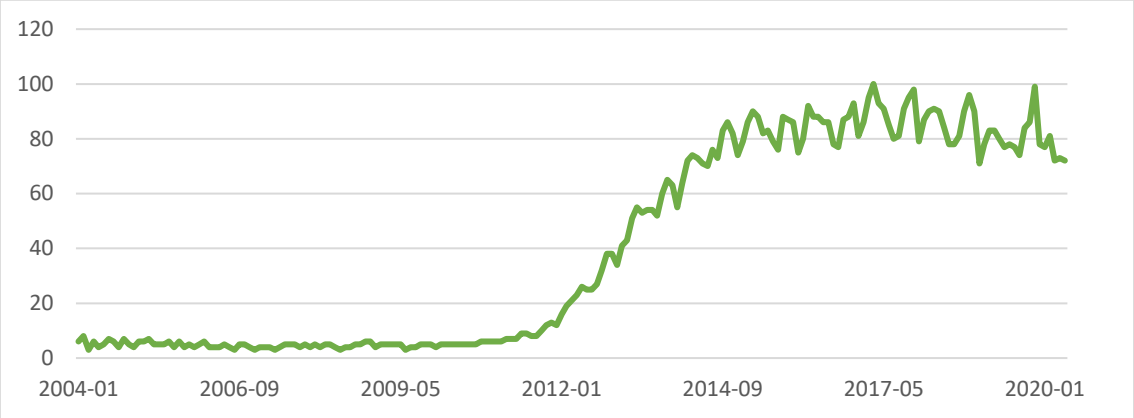


Figure 1 – Interest for the term Big Data, Google Trends (2020) (own elaboration)

According Jelonek (2017) based on Kraska (2013) and Fernández et al. (2014), the current explosion of data, that has led to the constant use of the term BD nowadays, is due to three main reasons:

1. A large amount of applications from sensors and social media services, among others, are continuously collecting information.
2. Technologies such as Cloud Computing make collecting data cheaper than ever, making its use affordable for any kind of organization, no matter the size or income.
3. Tools such as Machine Learning have reached a significant improvement and importance. This allows organizations the acquisition of a greater knowledge from the data collected.

This study aims to approach a topic, more often analysed and studied in computing sciences or mathematics, to the field of marketing, an interdisciplinary business area in charge of fulfilling customer needs (Szwacka-Mokrzycka, 2015). For this reason, the objective of this essay is to bring a reachable guide for directors of non-technical fields so that they can be aware of what is behind such powerful tools as the current ones for Textual Analysis, and how they can be a great boost to the company or organization.

The main contribution of this study is a description and summarization of few of the main technologies in Big Data. In addition, this research includes an introduction, explanation and the state-of-the-art of analytical tools currently used in Big Data context such as machine learning and Sentiment Analysis; tools used to improve the decision-making in the organizations.

This study is divided in two parts. In the first part, it is presented the Big Data scheme providing the reader with the basic background to better understand the technological context of BD and Sentiment Analysis. In this part, it is also described the business intelligence process matched with SA process in parallel. Then, the different technologies behind BD are described and contextualized as a tool to perform SA. These different technologies are introduced and classified within the different steps of the business intelligence process. The second part of the study consists in a systematic literature review. This will explore the literature that contains the terms (1) Sentiment Analysis, (2) Opinion Mining, (3) Marketing and (4) Machine Learning to analyse the literature so far and to glimpse the current problems of the organizations. Finally, future lines of research are presented to be considered for the extension of this study.

The following section will introduce what is behind the topic from a theoretical perspective, as well as its potential, describing the current situations organisations deal with.

Big data Ecosystem

What is Big data

It is not easy to categorize Big Data (BD). Nevertheless, for its characteristics, it can be considered as a topic within the area of Knowledge Management Systems. It is assumed to be part of this field as far as (1) it extends toward assimilation of information; (2) the role of Information Technologies (IT) involves collecting, storing and transferring knowledge; (3) it provides a connection between sources of knowledge to improve knowledge flows; (4) it involves effective retrieval mechanisms for pertinent information (Alavi & Leidner, 2001).

According Oxford English Dictionary, Big Data is "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data" (Simpson & Weiner, 1989, p1).

Nevertheless, the concept BD has been used along the years with different definitions. The first documented time the term BD appeared was in 1980 in a working paper from Charles Tilly that finally became a published article in 1984 titled 'The Old New Social History and the New Old Social History'. In this article, the author used the term as follows: "None of the big questions (in history) has actually yielded to the bludgeoning of the big-data people, and that in general the sophistication of the methodology has tended to exceed the reliability of the data" (Tilly, 1984, p. 369).

In the computing context, the term appeared in 1997 (Press, 2014). In an article titled 'Application-controlled demand paging for out-of-core visualization', the authors, both from MRJ/NASA Ames Research Center, mentioned that the large data sets (used in that time for visualization in the area of computational fluid dynamics) supposed a challenge in terms of the capacity of the computer's main memory to generate the appropriate visualizations in an effective way. The authors called this 'the problem of Big Data' (Cox and Ellsworth, 1997).

In 2001, Laney (2001) wrote a short article in which BD challenges were synthesized through the 3 Vs. These challenges were related within the data management context as a result of e-commerce; these are (1) Volume, (3) Velocity, (3) Variety. Since then, the 3Vs have been used to define or describe BD context (Press, 2014).

The awareness that BD era was a reality and has come to stay in the future, was revealed in an article in 2003 in which the author demonstrated how the generation of digital data had increased radically in the previous three years and showed how it would increase within the following years (Lyman, 2003).

Press (2014) also mentioned how the BD term had popularized after an article authored by three prominent computer scientists in 2008 was published. The article titled 'Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society' drew the context in which computational technologies such as (1) digital sensors, (2) computer networks, (3) data storage, (4) cluster computing systems, (5) cloud computing facilities and (6) data analysis algorithms; have changed the way in which companies, organizations, researchers, governments, among others, can access information (Bryant et al., 2008).

Knowing about where the term comes from is significant, but it is also important to acknowledge what kind of data BD involves. According Shim et al. (2015), based on an

article published by the consulting firm Gartner, there are five big groups of data which are captured by organizations. They are:

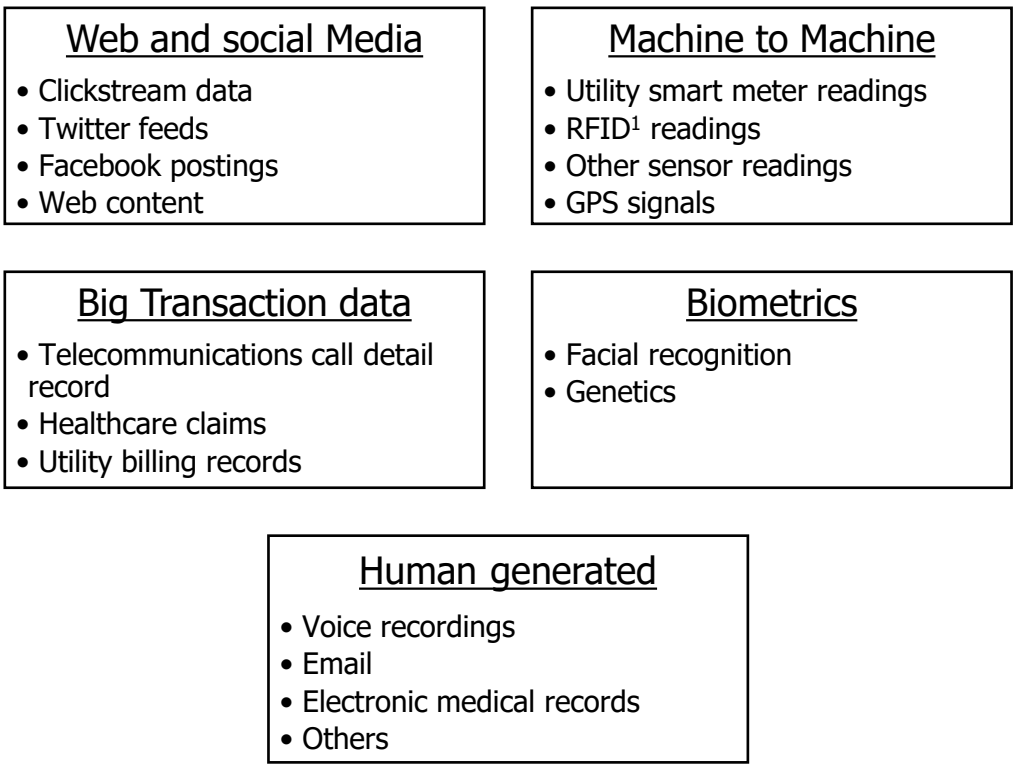


Figure 2. Five Big Data types. Shim et al. (2015) and (Mohanty et al., 2013). *(Own elaboration)*

[1] RDIF: Radio Frequency Identification

Regarding the industries, the use of BD and the area impacted is summarized in the following table:

Industry	**Impacted Area**
Manufacturing and logistics	Supply chain real time information
	Process analysis through sensors
	Logistics optimization
	Predictive maintenance
	Supply chain optimization
Retail	Cross selling (increasing the average purchase basket with similar products)
	Location-based marketing (to better target the consumer)

	Sentiment analysis (to analyse customer's response of the purchased items)
Telecommunications	Network optimization
	Churn prevention
Healthcare	Bioinformatics
	Simulation and prediction
	Tailor-made medicine
	Drug efficacy evaluation
Public sector	Boost productivity and efficiency
	Transparency

Table 1 Industries using Big Data (Serrato & Ramirez, 2017) *(own elaboration)*

Considering Laney's exposition about the Vs in BD, they can be described as follows:

Volume: Lyman highlighted how from 1999 to 2002 the digital information had doubled (Press, 2014). Moreover, considering the volume of data created worldwide from 2010 and the expected until 2025, these are the trends according Statista (2018):

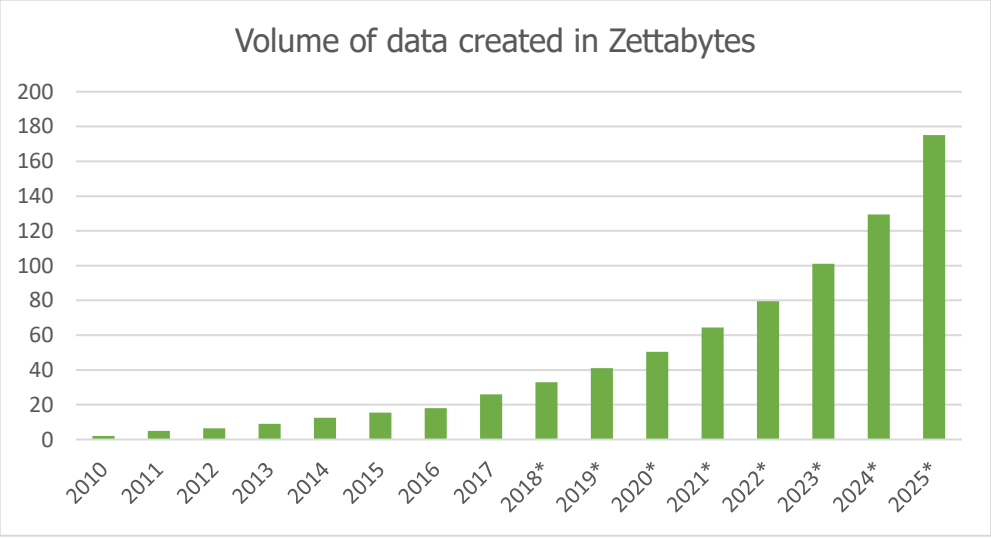


Figure 3 Volume of data created in Zettabytes[1] (Statista, 2018b) *(own elaboration)*

[1] Zettabyte is approximately equal to one thousand exabytes or one billion terabytes

Velocity: Refers to the speed in which data is generated and transferred. The new technologies have led to continuous flows of data at rates that have not been recorded before (Shim et al., 2015).

Variety: it is due to the heterogeneity in the data generated. In this context, it can be found mixed structured data, with unstructured and mixed data (Gandomi & Haider, 2015). As it will be seen later, Machine Learning can help to deal with unstructured data. In figures 2 and 4, the data generated comes from different sources. Therefore, the higher the variety so it is the complexity.

Veracity: Besides this 3 Vs previously described, some academics and companies such as IBM, included a 4th V for Veracity. This is the one related with the trustworthiness of the data. It represents the unreliability inherent to some types of data (Gandomi & Haider, 2015). The same authors pointed that veracity must be considered when conducting Sentiment analysis (SA) because of its uncertain nature (ambiguity).

In addition, other authors like Sivarajah et al. (2017) consider three added challenges to the previous four. These are:

Variability: it is due to the fact that in computational context the meaning of some data changes constantly. It happens for instance in SA where some words or sentences have a different meaning depending on the context or the way they are referred to (Sivarajah et al., 2017).

Value: it refers to the capacity and how valuable data can be to the organization or company (Shim et al., 2015).

Visualization: in order to take the right or accurate decisions based on data; it is important to be able to read and interpret data properly by getting spontaneous representations from the datasets. Tools such as Tableau or Power BI allow depictions for decision-making based on the data (Sivarajah et al., 2017).

The following figure summarizes how variety and complexity have increased in consequence of the data volume growth. The volume and complexity data growth in organizations have brought different technologies and contexts: (1) ERP (Enterprise Resource Planning tools); (2) CRM (Customer Relationship Management tools); (3) Web and (4) Big Data.

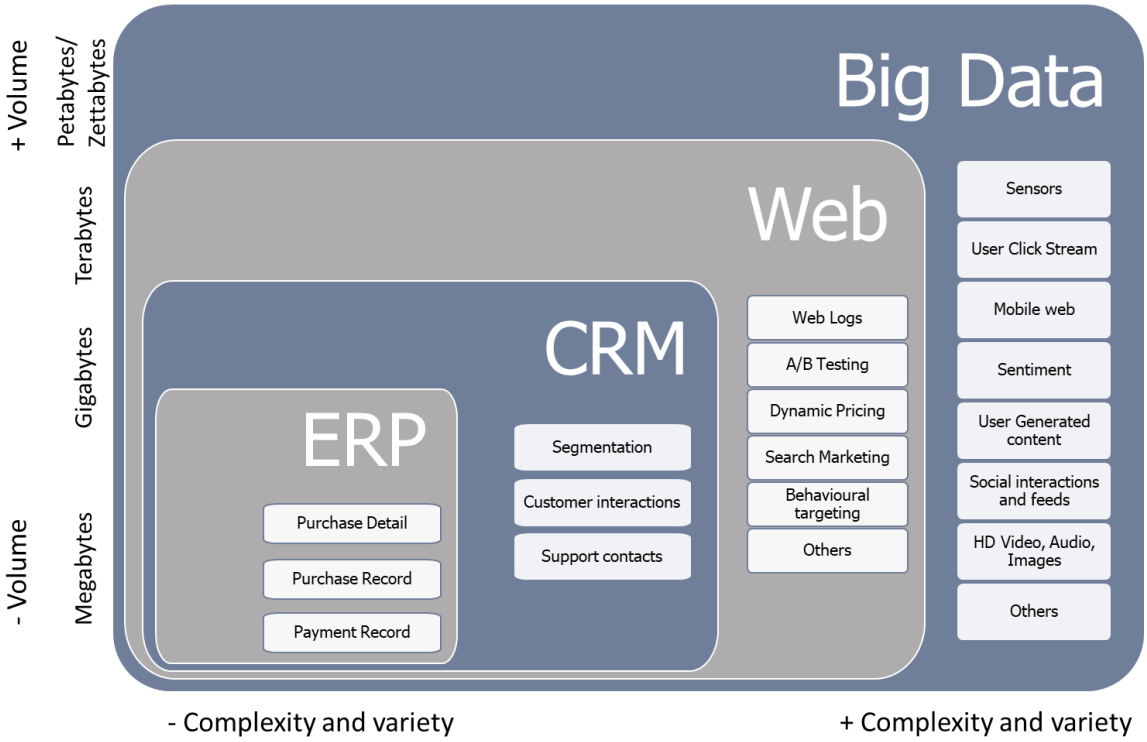


Figure 4. Big Data = Transactions + Interactions + Observations (Jelonek, 2017) (Own elaboration)

What has been described in this section are the context and challenges presented by the current situation of massive data generation that organizations deal with day by day. The way in which data is handled and managed, aimed to become information, can be a decisive factor in the survival and future of the organization. Next section introduces the different types of analytics to take advantage of the collected/obtained data.

From Big Data to Business Intelligence

As it can be deduced from the previous section, both complexity and opportunities have arisen within the context of Big Data. Nevertheless, as Powell and Dent-Micallef (1997) conclude in their article:

> "Information Technologies carry enormous productivity power but, like other powerful weapons, misfire in the wrong hands" (Powell & Dent-Micallef, 1997, p.396)

Business intelligence (BI) is defined by de Pablos, et al. (2019) based on article published by Forrester Research (2011) as the design and implementation of an architecture or infrastructure as well as processes and best practices for the correct (1) storage, (2) integration, (3) communication and (4) analysis of business information. The same authors highlighted that with BI it is aimed to bring an integral management of the

information within the organization in order to get a better competitiveness (de Pablos et al., 2019).

Regarding 'analysis', an important step to extract the real value of the data, there are three main ways in which organizations can manage Big Data:

1. Descriptive analytics: this is the most common type of analytics; it answers the question 'what has happened' referring to the past. It uses descriptive statistics, and basically monitors organization's data and it is in charge to detect deviations from previously defined Indicators (KPIs) (Appelbaum et al., 2017).
2. Predictive analytics: a further step from the previous one. It answers to question 'what could happen'. It uses probability models to calculate and make predictions for possible future events. In addition, Predictive analytics attempts to establish relationships between variables through mathematical methods to draw conclusions (Waller & Fawcett, 2013).
3. Prescriptive analytics: this type of analytics goes beyond descriptive and predictive analytics and answers the question 'what should be done considering descriptive and predictive output'. This kind of analytics can improve the accuracy of predictions and consequently contributes to a better decision-making considering different scenarios (Appelbaum et al., 2017).

This study aims to introduce the context and the concept behind Sentiment Analysis for a better decision-making within the organizations, taking advantage of the content generated by humans, concretely product reviews. Previous studies like the one conducted by Appelbaum et al. (2017) affirmed that the three types of analytics previously described are useful to better understand Customer perspective.

This section has introduced briefly the process or path from Big Data to Business Intelligence (Value). Combining de Pablos et al. (2019) and Sivarajah et al. (2017), the process can be synthetized as follows:

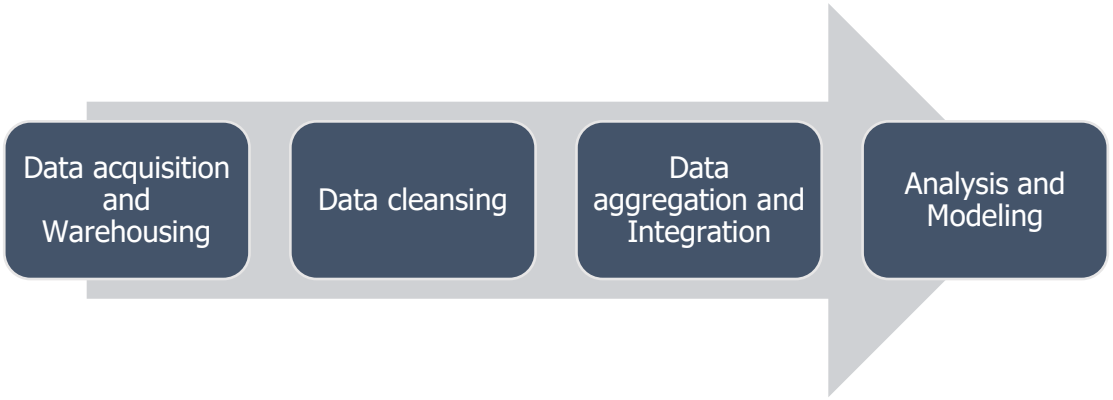


Figure 5 Business Intelligence process (Sivarajah et al., 2017) *and* (de Pablos et al., 2019) *(own elaboration)*

The steps presented in Figure 5 will be detailed and described together with the technologies involved in later sections.

As seen in figure 5, the starting point is gathering data. The difficulty in this first step is that the type of data used to conduct analytics from customers comes both in a structured and unstructured way (Variety). This fact can bring the organizations to biased or wrong conclusions.

Next section introduces the theoretical framework of Sentiment Analysis and Text analytics and how it can be performed properly to add value in the organization.

Sentiment Analysis

Sentiment Analysis (SA) or Opinion Mining can be described as a set of techniques used to analyse opinionated text that contains people's opinion towards different entities such as products, services, organizations or individuals, among others (Gandomi & Haider, 2015).

Textual data on the Internet is growing at a rapid pace and many companies and organizations are attempting to use this data stream to extract people's point of view regarding their products (Sheela, 2016).

Notably both SA and Business Intelligence needed to follow a process to extract proper conclusions from the data. Therefore, the BI process, as shown in figure 5, is considered to be the model for SA in this study.

While analysing the literature, different terms connected to SA have been detected. For this reason, they should be defined first, in order to better understand what underlies behind SA process:

Text analytics (TA) can be defined as the techniques used to gather information from textual data generated for users including product opinions, social network post, online forums, emails, blogs, responses from surveys, reports, among others. In addition, TA enables organizations to effectively handle and manage large volumes of human generated content that are susceptible to become an insight and high valuable information for the organization (Gandomi & Haider, 2015). The same authors equate the term 'Text Analytics' with 'Text Mining'.

Text Mining (TM) involves statistical analysis, computational linguistics and machine learning, among others (Gandomi & Haider, 2015). According Hotho et al. (2005), TM can have a different meaning depending on the research area. They detailed these:

1. Text Mining as Information extraction; it refers to the extraction of facts from text.
2. Text Mining as Text Data mining; this implies algorithms and methods from Machine Learning and statistics that aim to find patterns in the data.
3. Text Mining as Knowledge Discovery and Data Mining; this involves how data is (1) stored and accessed, (2) the developing of efficient and scalable algorithms, (3) the interpretation and visualization of the results, and (4) the modelling and interaction between human and machine (Kurgan & Musilek, 2006).

Consumer analytics in Big Data context, it can be defined as the extraction of consumers' unperceived insights from the complexity of Big Data, exploring them through a profitable interpretation (Erevelles et al., 2016).

Data mining (DM) is defined by Ertel (2017) as "the task of a learning machine to extract knowledge from training data" (p.179). The same author also described DM as the knowledge acquisition process from data through statistics or ML in a context of large volumes of data at reasonable cost (Ertel, 2017).

Sentiment Analysis can be applied in different areas: (1) commercial product area, (2) politics area (citizenships opinion regarding certain topics and political elections, among others), and (3) stock market forecasting (Hemmatian & Sohrabi, 2017). This study focuses on the first one. According to the same authors the achievement expectations through SA in the commercial product area are:

- Product comparison
- Sentiment summarization
- Exploring the reason of opinion

Next section will introduce the technologies involved through the different steps of the process.

Big Data techniques

Serrato and Ramirez (2017) exemplified in an article the case of eBay, which used in 2016 (when the article was written) two data warehouses at 7,5 petabytes (7500 terabytes) and 40 petabytes Hadoop cluster (a tool which will be analyzed below) for searching and analyzing consumer recommendations. These amount was stored across company's distributed cluster and was managed with Hadoop, machine learning and cognitive computing (Serrato & Ramirez, 2017).

A detailed analysis of all the technologies involved in SA, which are part of BD, is not the purpose of this study; as it is not possible due to the large number of technologies involved which at the same time evolve day by day. What is intended to do in this section is presenting and briefly clarifying some of the tools and techniques that make SA possible nowadays. These will be presented within the different steps of BI model (figure 5).

Data acquisition and Warehousing

Data acquisition

As discussed in the previous section, for the last few years and currently the volume of data generated has been increasing radically. This vast amount of data can be obtained from different sources (Figure 2). Nevertheless, this study is focused in the data generated in internet by users in e-commerce portals.

There are different ways to obtain the data from websites but a technique that proves to be very efficient is **web scraping**. This technique is employed to retrieve data from websites using scripts named crawlers. One of the advantages of Web scraping is that once the code is created and executed, data can be extracted from the defined domains (websites) automatically (Prathi et al., 2020).

There are several means to implement web scraping. However, even though this study will not focus on them, it is important to remark that the mastering of this technique can play a key role for the success of the BI process.

Data warehousing

The vast amount of data generated day by day and which can be highly valuable, it is hardly supported by a single machine. Current technologies have been evolved while complexity grew to provide solutions to the one of big challenges of Big Data, Volume. One of these technologies, of which one of its main functions is the correct warehousing and data handling, is Cloud Computing (CC).

However, before defining this technology it is important to enumerate the different types of storage for Big Data. These are:

Storage type	Solution
Block-Based	Amazon EBS: it is one of the most used solutions since it can handle intensive workloads at any scale (high volume of data).
	OpenStack Cinder: open source cloud solution similar to Amazon EBS.
File-Based	NFS family (Network File System): is a protocol in which files can be accessed as if they were on the local machine, even though they are stored in a remote machine.
	HDFS (Hadoop Distributed File Systems): it is a distributed file system that supplies scalable and reliable data storage over a large set of server nodes.
Object -Based	Amazon S3: it stores data as arbitrary objects, organized into buckets. It allows data/objects manipulation (creation, retrieving, listing). It is a scalable solution with a millisecond latency[2].
	OpenStack Swift: as Amazon S3 it is a highly scalable and low latency solution which uses software logic to secure data replication data replication over different devices.

Table 2 Storage Models (Zomaya & Sakr, 2017) (Own elaboration)

[2] Latency: the delay before data begins to move after it has been sent an instruction to do so (Oxford dictionary, 2020)

All these types of data storage are available and are an option in the wide range of solutions that CC technologies offer. Currently, the types of storage are the ones shown in Table 2 ("File storage, block storage, or object storage?," n.d.); nevertheless, regarding the solutions by type of storage, it cannot be accurately stated if there are more of the ones presented since the literature consulted cannot confirm it. Following, it is defined this technology behind.

Cloud

Cloud Computing or Cloud technology is a technology that comes as a part of the solution for the challenges mentioned in the previous section. CC can be consider as an extension of Information and Communication Technologies (ICTs) (Candel Haug et al., 2016). According to the same authors and based on the theory argued by (Henderson & Clark (1990), the cloud emerged as an architectural innovation as a result of isolated innovative processes in order to transmit data agilely.

In addition, Cloud computing is an internet-based technology, through which information is stored on servers and provided as a pay-per-use service (Sánchez & Correia, 2016). The cloud allows access to companies and end-users to updated software and infrastructure solutions, among others, at low cost; access to their data from anywhere and at any time allowing the flexibility and portability that users demands nowadays (Assante et al., 2016). Moreover, the fact that the server is remote provides advantages to the company such as lower cost of maintenance and savings in physical space (Simalango et al., 2010).

Originally, Oliveira et al. (2014) exposed that CC includes three main areas which include:

- IaaS (Infrastructure-as-a-Service): which includes the infrastructure for storage according to demand. The main suppliers are companies such as Rackspace or Amazon Elastic Compute Cloud.
- PaaS (Platform-as-a-Service): in this case what is offered are integrated solutions for creating and launching applications from the cloud. Among the suppliers are Microsoft Azure and Google AppEngine.
- SaaS (Software-as-a-Service): allows access to end users to applications hosted centrally in the cloud through browsers. Example of companies that offer this service are Salesforce or GotoMeeting.

Regarding the shares of the different pay-per-use services; SaaS is the Cloud service more expanded with 71.8% of use respect IaaS which has 19.4% of coverage and PaaS with 8.8% (Statista, 2018a).

Besides the three main pay-per-use services, there are still three more not so well known. These are:

- FaaS (Function-as-a-Service): it facilitates serverless computing by introducing a cost-efficiency solution for the client by reducing configuration and management overheads. Lambda from Amazon Web Services (AWS) is an example of FaaS (Lynn et al., 2017).
- DaaS (Data-as-a-Service): this solution provides the client the possibility to download data form cloud providers ensuring (1) speed, (2) cost effectiveness and (3) reliability (Al Nuaimi et al., 2015). Database tools such CassandraDB or business solutions like ERPs or CRMs (Melgrati, 2018).
- STaaS: (Storage-as-a-Service): it is a service in which the organization, company or individual just rents the storage infrastructure for its files. It can be found examples of STaaS in cloud storage solutions like Dropbox AWS S3, OneDrive, among others (Melgrati, 2018).

In summary, Cloud Computing technologies offer a wide range of solutions for companies and organizations based on remote servers. These solutions can be summarized as (1) Storage Systems, (2) Processing Systems, and (3) Communications systems. The main providers currently of these infrastructures are Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform (Vivancos, 2018).

Data Cleansing

This phase of the BI process ensures the veracity of the collected data. It addresses the problem of missing, redundant and uninformative data. It also tackles the issue of format and compatibility between different sources (Curuksu, 2019). The author also pointed that this phase precedes one of the most critical phases, Data integration from different sources.

In the context of Sentiment Analysis, data cleansing is associated with the data pre-processing requirement. This step involves handling the inherent noise of human generated text (Karacapilidis et al., 2013). This will be explained with more detail in the phase Data Analysis and Modelling.

Data Aggregation and Integration

As mentioned in the previous section, data comes from different sources that have their own architecture and format. This fact supposes an important challenge that organizations must be aware to select the option and tools that suit them best. The main aim of this process is that all data sources gathered from different sources can efficiently answer queries that embrace different sources. Other objective is the design of a scheme that allows the aggregation of new data easily (Doan et al., 2012).

Cloud technologies offer different kind of solutions to pool the data depending on the size space requirement. these are summarized in the following figure:

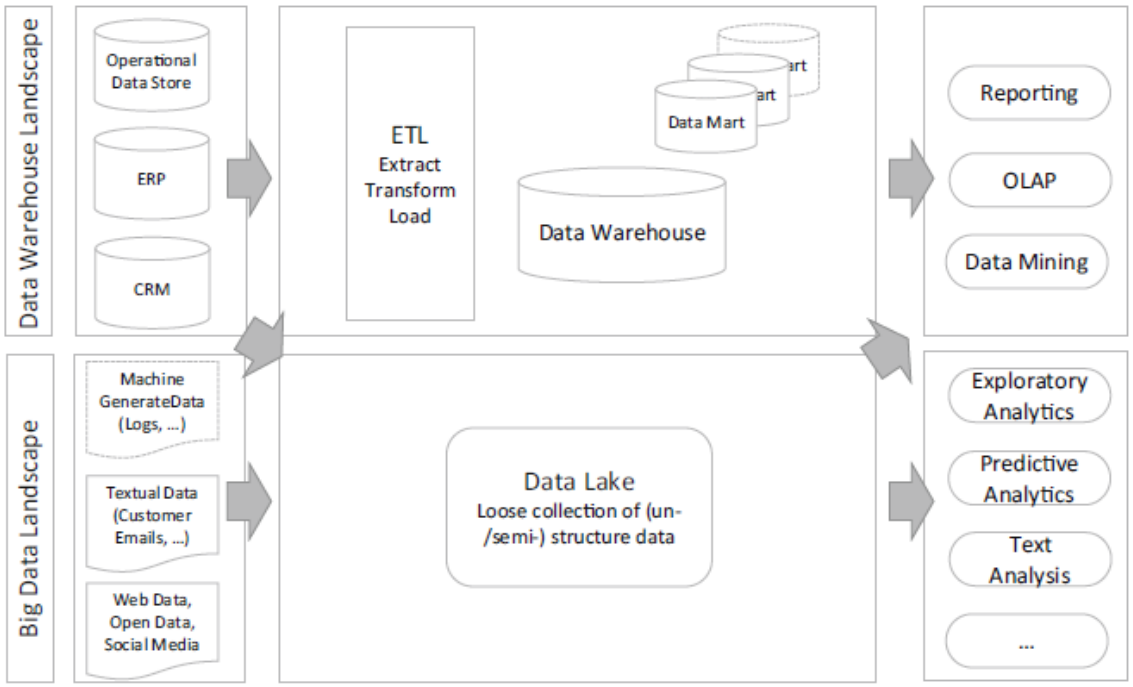


Figure 6 The growing big data analytics landscape (Eberius et al., 2017, p. 366).

As seen in figure 6, Big Data Landscape approaches to the aim of the study in a better way. Therefore, this phase considers Data Lake as the proper approach to handle the data from different sources.

The concept Data Lake comes from James Dixon; he promotes this concept as a new way of managing big data that come from wide world connected devices (Woods, 2011). One of the attributes of Data Lakes is that it keeps the data in a raw format; and it is later transformed (through a raw copy) to be analysed. So, what Data Lake stands for is to preserve value from data that comes in a (1) structured, (2) semi-structured, and

(3) unstructured way making it accessible whenever necessary. This agility in data accessibility occurs through queries in the Data Lake platform (Woods, 2011).

During the last years, Hadoop has been one of the preferred options for building Data Lakes. This is due because it a highly scalable framework that enables the processing of large datasets. Hadoop is an Open Source Solution which includes a software framework to handle the challenges of volume and variety and veracity. Broadly, Hadoop consists in "first mapping the data and then reducing it, finally and based on nodes distributing tasks in a large network of servers" (Shim et al., 2015, p.802).

There are further solutions beyond Hadoop, it will be listed some of them below these lines. Nevertheless, the analysis in deep of these, including Hadoop, are not the aim of this study as it corresponds to more specialized fields like Computing engineering or Telecommunications engineering, which are able to provide a more accurate and proper definition, description as well as applications.

Apache Spark
Apache Storm
Ceph
Hydra
Google BigQuery

Table 3 Hadoop alternatives (Tutorial, 2020)

What has been presented briefly in this subsection is part of the context in which the aggregation and integration of Big Data is managed and handled nowadays. The aim is to familiarize a non-technical reader with terms and procedures that can be part of SA.

The following phase moves from Big data to the Artificial intelligence (AI) paradigm through Machine Learning.

Analysis and Modelling

This is the most decisive phase of the process. To manage it rightly, it is needed first that previous phases have been conducted correctly. The purpose of this phase is to stablish relationships and patterns between the data obtained, stored and integrated (Sivarajah et al., 2017). As BD management carries complexity, approaches that come

from Artificial intelligence area can help to stablish more accurate relationships for a better output considering future scenarios (predictive analysis).

Artificial intelligence first appeared in the academic field with the name of Machine Intelligence from Alan Turing in 1950 in an article titled 'Computing Machinery and Intelligence' (Ertel, 2017). Since then, the term and its background has evolved. According Investopedia, "AI refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions" (Scott, 2020, p.1). To achieve human intelligence simulation, AI includes these following subareas:

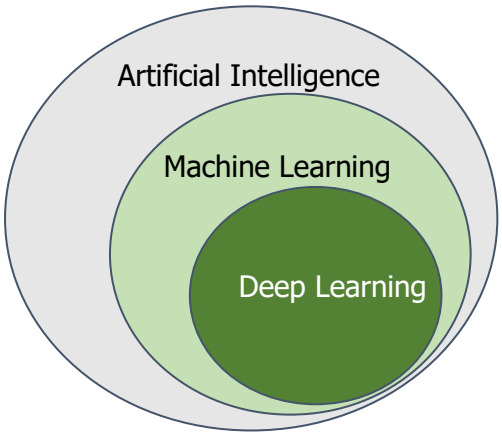


Figure 7 Artificial Intelligence subareas based on Patterson & Gibson (2017) *(own elaboration)*

As it can be seen from figure 7, Machine Learning (ML) is a subarea of AI. This study focuses on Machine Learning as the first approach to be part of the BI process.

Nevertheless, one of the questions organizations should consider is if they are ready to implement AI systems such as ML. According Earley and Bernoff (2020), it is necessary for organizations to build an ontology considering these steps: (1) Identification of data pain points, (2) generation of solutions based on root causes, (3) understanding the use of cases, and (4) setting the ontology organization principles.

Machine Learning

ML is a subarea of AI that involves self-learning algorithms that derive knowledge from data to make predictions (Raschka, 2015). In addition, ML algorithms can learn to perform tasks without specific instructions, relying on patterns discovered in data. For instance, algorithms can be performed to predict the likelihood of having a disease or assess the risk of failure in complex manufacturing equipment (Dubovikov, 2019).

ML applied in in the field of Sentiment Analysis can be useful for organizations as far as it (1) provides clients, manufacturers and other stakeholders information about the characteristics or features more and less appreciated by the final customer; (2) allows the detection of the competing products and how customer perceives them in comparison with the owned by the owned products/services (Shafaee et al., 2014).

There are different aspects to consider when explaining ML. One of the main aspects is the 'why' of this tool. ML can solve three types of problems (Henke et al., 2016):

1. Classification: by identifying objects and recognizing text or audio. Classification also includes associations and recommendations through segmentation into clusters.
2. Prediction/Estimation: to predict and forecast possible outcomes
3. Generation: in this case for instance ML "can generate content from interpolating missing data to generating the next frame in a video sequence" (Henke et al., 2016, p.12).

This study and the field of Sentiment Analysis focuses on Classification and Prediction. ML is a technique that has evolved and has 3 main different types of learning depending on the data and purpose:

Branch	Types of data	Description	Form of analysis
Supervised learning (SL)	Labelled data	Through SL workflow the labelled training data is analysed through a ML algorithm to define a predictive model; thus, it can predict new unlabelled data inputs (Raschka, 2015).	Regression Classification
Un-supervised learning	Unlabelled data	The purpose is uncovering hidden structure data through unlabelled data (without historical labels) to train the algorithm. (Ramzan et al., 2019).	Clustering Dimensionality reduction
Reinforcement learning	System Agent	It consists in an interaction with the environment in which the agent through a trial-error process can reinforce the learning by maximizing its reward (hit accomplishment) (Raschka, 2015).	Reward System

Table 4 Machine Learning branches (Ramzan et al., 2019) *and* (Raschka, 2015) *(own elaboration)*

In addition, there are several terms that should be taken into consideration to understand the ML process:

- Accuracy: in the ML context, it is defined by Subroto and Apriyana (2019) as the primary measurement used for model selection through the following formula:

$$Accuracy = \frac{(True\ Positives + True\ Negatives)}{(True\ Positives + False\ Positives + True\ Negatives + False\ Negatives}$$

- Training: "is the process of taking content that is known to belong to specified classes and creating a classifier on the basis of that known content" (Marklogic Coorporation, 2019, p. 726). There is a particular body of sentiment datasets which are available for training the data (McCartney & Potts, 2020):
    - Amazon Customers Review data
    - Amazon Product data
    - Stanford Sentiment treebank
    - IMDb (movie dataset)
    - Others.
- Test: it is used to evaluate whether the trained agent can be extrapolated from the training data (Ertel, 2017).

The following Figure shows the steps for conducting a ML model:

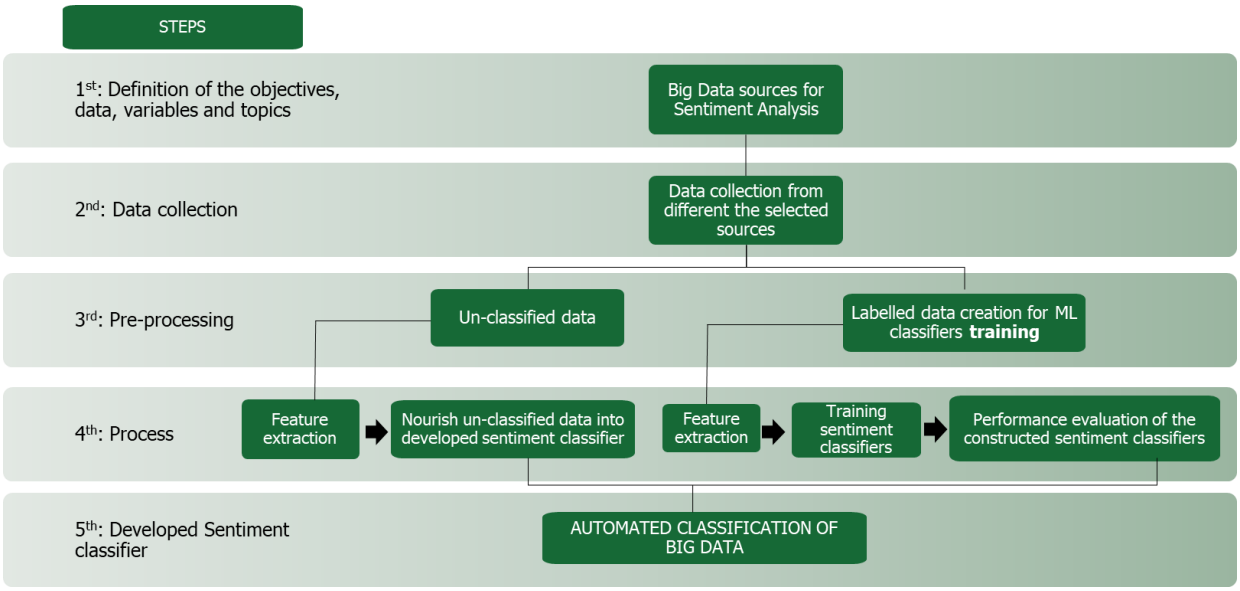


Figure 8 Steps for Machine Learning method construction (Shayaa et al., 2018) *(own elaboration)*

Approaches such as 'Extraction based on topic modelling' can be a good option for Sentiment Analysis to detect hidden topics in the text that are not labelled (un-supervised learning). The aim of this approach is automatically obtaining relevant insights from data without human supervision (Hemmatian & Sohrabi, 2017).

Further approaches to handle Big Data

As seen in the previous section, the Business Intelligence process, considering Big Data context using tools like Machine Learning, is complex to execute. For this reason, including approaches such as Fuzzy logic can ease the extraction of value from data.

Fuzzy logic is defined as "a system of logic in which a statement can be true, false or any of a continuum of values in between" (Merriam-Webster, n.d., p. 01). Fuzzy logic can be implemented in contexts that involve complexity issues and vagueness as it allows more flexible mathematical schemas closer to the reality (Kaufmann & Gil-Aluja, 1986). Fuzzy logic is based in the principle of gradual simultaneity in which any proposition can be considered either true or false by allocating a degree of trueness and falseness (Gil-Lafuente et al., 2017).

The Fuzzy Sets method can be used to analyse social systems for their capacity to deal with (1) vagueness, (2) ambiguity and (3) uncertainty (veracity challenge) of qualitative data derived from judgements; this method can be used to model raw sentiment with classification probabilities through α cut technique (Mukkamala et al., 2014).

Ontology in information science context can be defined as a set of attributes, relationships and principles that provide a common understanding of a concrete context through disambiguation. This allows users and systems to stablish a type of communication each other along a simpler information exchange and better integration (Calegari & Ciucci, 2006). Nevertheless, when the degree of complexity is high, ontology for itself may not be enough. For this reason, the incorporation of fuzzy set theory to ontology can be a solution to deal with the analysis of semantic websites such as e-commerce or other websites (Calegari & Ciucci, 2006).

Lexicon-based approach is a method that gives specific weights for each word of the text considering its belonging polarity (negative, positive or neutral). This methodology can be performed by using resources such as SentiWordNet or Vader, among others (Al-Shabi, 2020). For its characteristics this approach can be performed combined with fuzzy Logic.

Moreover, for Opinion mining, a subfield of textual analysis, it should be considered three aspects: the analysis through a (1) lexical, (2) syntax and (3) semantic perspective to understand sentiment towards specific features (Ramzan et al., 2019). Some articles have considered Fuzzy Logic domain combined with Ontology to extract more accurate conclusions by considering not only the lexical and the syntax aspect but semantics,

which approximates human languages to computational context. This combination has been used to analyse Recommendation systems based on the cases of guest's opinion mining with regard the hotel they stayed (Ali et al., 2016), (Ramzan et al., 2019) and (Calegari & Ciucci, 2006).

It can be concluded from this part that Big Data brings challenges that need to be considered but also provides solutions that can suppose an opportunity for those organizations able to handle and take advantage from it. The defined and described concepts from this section will help to better understand the following section.

Methodology

The methodology of this study is a Literature Review (LR) of the topic Sentiment analysis and Opinion mining. Conducting this LR, it is aimed to be able to answer the following questions:

- **Q1:** Does academic marketing journals provide articles related to Big data and Sentiment Analysis?
- **Q2:** Which are the current trends in Sentiment analysis/Opinion Mining research and the most reliable tools in Machine Learning?
- **Q3:** Is ML and SA reliable for predictive and prescriptive analytics for e-commerce businesses?

As a first approach, it has been carried out an analysis through Journal Citation Report (JCR) looking for which are the main 'business' journals that publish marketing articles.

Once the list of journals has been obtained, it has been checked if these considered BD as a topic in their publications; and if so, how many articles have included the topic.

It has been conducted a search considering the categories (1) Business and (2) Management in the first JIF[3] Quartile, the list obtained has included 70 journals. From this selection, the next step has been selecting the journals ranked as the ones more cited. Finally, the chosen journals have been the 'Journal of Marketing' and the 'Journal of Marketing Research'.

[3] Journal Impact Factor

Through these two journals, it has been looked for articles from 2000 using the key words (1) Opinion Mining and (2) Sentiment Analysis. The articles obtained are as follows:

Journal of Marketing[4]	11 in total. Only 2 of them related with this research	More than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates (2013)
		Uniting the Tribes: Using Text for Marketing Insight (2019)
Journal of Marketing Research	5 in total. Only 2 of them related with this research	Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation (2014)
		Automated Marketing Research Using Online Customer Reviews (2011)

Table 5 Interdisciplinary journals (own elaboration)

What can be concluded from this first approach is that notable journals, such as the previously analysed in table 5, specialized in the area of marketing, do not include in its issues many articles related with Opinion Mining. Considering these limited first results, it has been decided to conduct a more extensive LR through Web of Science website.

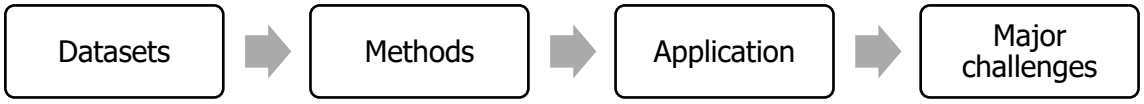
In an article authored by Shayaa et al. (2018) it was conducted an accurate LR. In their article, the authors started the research by looking for the following terms through Web of Science, searching 'Opinion mining' and 'Sentiment analysis', using Boolean logic operators. The results they obtained as well as the research criteria are as follows:

Criteria	Exclusion criteria
• Articles published in peer-reviewed journals • From 2001 to 2016 • In English	• Articles with a narrow context • Not explicitly focused on the application considering human or organizational level
The result obtained applying these criteria was 365 articles. From this result the authors selected the 99 considering the Research questions exposed in their work	

Table 6 Article selection criteria carried out by Shayaa et al. (2018) (own elaboration)

[4] From Sage journals

In addition, Shayaa et al. (2018) propose a taxonomy to analyse the selected articles. It is composed by the following steps:



In this study, the starting point is similar as Shayaa et al. but in this case, it has been decided to add the keywords 'Marketing' and 'Machine Learning' besides 'Opinion Mining' and 'Sentiment Analysis'.

Doing so, and applying the following filters:

(1) Language: English

(2) Timespan: from 2016 to 2020

(3) Document type: article

The number of articles obtained is 79. From these articles, a second selection has been made excluding those papers related exclusively with Social Media platforms, exclusively from technical or linguistics fields and those with datasets not related with e-commerce (energy, financial, tourism). As a result, the final list of selected papers to be analysed is 26.

Then, it has been considered to include the following variables to be able to give an answer to the research questions:

- Year
- Journal
- Dataset and/or data sources
- The objective of the study
- The main used techniques
- The results

Paper	Year	Journal	Dataset/ Data sources	Purpose/objective	Main techniques	Results
[1]	2020	Expert Systems with Applications	Bing Liu's dataset[5]	To introduce two lexicon generation methods to adequately adapt general lexicons to the context.	(1) Aspect-Based Frequency Based Sentiment Analysis (ABFBSA) lexicon generation	The proposed approach outperformed baseline methods in aspect-based polarity classification by improving the accuracy.
[2]	2020	Future Generation Computer Systems	Set of reviews from Apps in Google play	To propose a methodology to automatically extract the features of an app from client's reviews.	(1) Formal Concept Analysis (FCA) Algorithm (2) Weighted-Tree Similarity Algorithm	The authors developed a hot feature extraction algorithm which demonstrates tree-based techniques as a good tool for precision and recall.
[3]	2020	International Journal of Intelligent Systems	Reviews from tmall.com jd.com and suning.com	To present an improved method for sentiment orientation calculation of multidimensional attributes based on product attribute classification framework.	(1) Deep-Learning (DL) based opinion (2) *q-rung* orthopair fuzzy interaction weighted Heronian mean (q -ROFIWHM) operators	The algorithm proposed can automatically extract product attributes from reviews in a wide range of adaptabilities (different e-commerce platforms).
[4]	2020	Journal of Cloud Computing	Movie Reviews dataset from Pang et Lee	To provide a solution to analyse and classify sentiments into positive and negative classes by identifying the polarity of words.	Deep Learning algorithms: (1) Convolutional Neural Network (CNN) (2) Long Short-Term memory (LSTM)	Their implementation showed that their proposed model has an accuracy of 89.02%

[5] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets

[5]	2020	Information Processing and Management	Stanford Sentiment dataset	To propose a distributed Intelligent system to handle real-time stream processing.	(1) LSTM (2) Bi-directional LSTM (3) Gated Recurrent Unit (GRU)	The results show that the three presented techniques outperformed Recurrent Neural Network (RNN).
[6]	2019	Engineering Applications of Artificial Intelligence	Product reviews in Amazon.com	To present an alternative to Kansei engineering based on a heuristic deep learning method.	Classification method trough rule-based extraction and Deep learning, combined with: (1) K-nearest neighbour (KNN) (2) Classification and Regression Tree (CART) (3) Multi-layer perception (MLP) (4) LSTM (5) Deep Belief Network (DBN)	the proposed method has an accuracy of 86% outperforming baseline methods.
[7]	2019	Decision Support Systems	Reviews from autohome.com.cn and bitauto.com NTUSD (National Taiwan University Sentiment Dictionary)	To introduce a method based on User-generated content (UGC) that reveals advantages and disadvantages of a target product compared with its competitors.	5 standard classification methods: (1) Decision tree (2) Decision table (3) KNN (4) Logistic regression (5) Naïve Bayes	A useful method that from supervised learning identifies competitors from UGD analysing and quantifying customer attitudes.
[8]	2019	International Journal of Decision Support System Technology	Movie Reviews dataset from Pang et Lee	To evaluate the performance of some Machine Learning techniques.	(1) Maximum Entropy (2) Naïve Bayes (3) Support Vector Machine (SVM)	Sentiment Analysis using SVM outperforms other machine learning techniques

[9]	2019	Electronic Commerce Research and Applications	Amazon Product Data	To present a unified approach for learning to rank the products based on online reviews.	(1) Hierarchical Attention Network for learning to rank (HAN-LTR) (2) CNN (3) LSTM	The authors demonstrated how HAN-LTR outperformed other methods for sales rank prediction considering online reviews
[10]	2019	Computer Speech and Language	IMDb movie dataset Czech-Slovak Film Database (CSFD) Amazon reviews from different locations	To address the challenge of SA considering multiple languages and different thematic domains though four examples by incorporating the surrounding context.	(1) SVM (2) n-gram based model (3) Term Frequency-Inverse Document Frequency (TF-IDF)	The authors proved that a classifier that considers labelled data from multiple languages is effective and performs suitably. They suggested also to incorporate surrounding context in future research.
[11]	2018	International Journal of Production Research	Amazon reviews WordNet database	To propose an approach based on semantic orientation to measure through a redesign index the priority of product feature redesign.	(1) Hierarchical clustering approach	The results demonstrate that the proposed approach provides an effective way to identify the product features that need to be enhanced.
[12]	2018	Journal of Intelligent & Fuzzy Systems	Amazon reviews, Flipkart and Snapdeal using the web Scraping platform import.io	To suggest an approach that combines structural and content-based features from reviews to rank online products through Opinion-based Multi-criteria Ranking (OMCR).	(1) Analytic Hierarchy Process (AHP) (2) Set intersection method	OMCR approach provides an evaluation results of 83.67% for smartphones and 84.33% for hard disk drive.

[13]	2018	IEEE Access	Amazon reviews	Investigate trust between users in e-commerce systems in a quantitative way evaluating their opinions towards commodities, services and business among other subjects.	(1) TF-IDF	The authors also concluded that user's trust relationship can be obtained regarding the similarity of them.
[14]	2017	Expert Systems with Applications	Product reviews in Amazon.com Wordnet database	To incorporate client's preference information into feature models using SA from user's product reviews.	Hybrid SA combining lexicon-based and ML methods: (1) Applied Association Rule Mining (ARM) (2) Rough set technique	The authors concluded that the proposed method can enhance the quality of product line planning considering customer needs.
[15]	2017	International Journal on Artificial Intelligence Tools	Yelp Review dataset	To provide a system that firstly allows user to train models for performing Aspect-level SA tasks; and finally, a web application in which the end-user can introduce reviews to analyse Aspect-level sentiments.	(1) SVM (2) Random Forest (3) XGBoost (4) Naïve Bayes	An open source system for Supervised ML to extract aspect terms and predict sentiment labels.
[16]	2017	Information Systems Frontiers	IMDb Amazon review datasets Twitter	To present different ML techniques demonstrating which is better according the dataset used.	Classification ML techniques: (1) SVM; (2) Naïve Bayes (3) Decision tree	The experiments show that there is sensitivity between the ML technique and the dataset properties because of the size, length, subjectivity, among others.

[17]	2016	World Wide Web	Epinions Flixter Ciao	To propose a method that uses Similarity-based Sparsification techniques to uncover the edge types connected to the nodes in the different user's communities.	(1) graph clustering (2) Localized Community Detection Algorithm (LCD)	Through LCD the authors have grouped successfully similar reviewers to improve filtering according user's personalized information.
[18]	2016	Expert Systems with Applications	Stanford Sentiment dataset Natural Language Tool Kit (NTLK) IMBd Movie Review Data from Cornell	To propose a methodology for sentiment detection through ML considering different languages.	(1) Hybrid vectorization (2) SVM	In all the experiments performed, hybrid vectors provide a more accurate result. It is also concluded that the methodology proposed can be applied for any language with similar characteristics.
[19]	2016	Information Processing and Management	Amazon reviews Bing Liu's dataset	To investigate the combined effect of ML classifiers and sampling methods in sentiment classification with unbalanced data.	(1) SVM based ensemble algorithm (2) Operating Receiver Characteristic Curve (ROC)	The ensemble SVM applied showed improvement in prediction performance, not only for majority class but for minority class.
[20]	2016	International Journal of Information Technology & Decision Making	Amazon's, Ebay's, Cnet's and Epinion's reviews SentiWordNet	To propose a model that includes aggregation methods to evaluate a product based on its features.	Aggregation methods: (1) Ratio of Positive over All Opinions (ROPOAO) (2) Difference of Positives and Negatives over all opinions (DPNOAO)	The results show that the proposed model is robust to detect false ratings and provides a good estimation reputation value based on the reviews.

[21]	2016	Cognitive Computation	Amazon reviews NTLK Jieba (Chinese NTLK version)	To address the problem of studying consumer behaviour taking advantage of opinion mining as a substitute of questionnaires to analyse difference between Chinese and American customers.	It is performed multi-granularity SA: (1) Latent Dirichlet Allocation (LDA)	The authors found several differences between the two markets. They also highlighted that generate questionnaires automatically integrating opinion mining is a field for further research.
[22]	2016	Applied Soft Computing	Amazon and e-Bay reviews Stanford datasets	To introduce a new computational intelligence framework for customer review ratings prediction.	(1) Singular Value Decomposition (SVD) (2) Fuzzy C-Means (3) Adaptative Neuro-Fuzzy Inference System (ANFIS)	The performance of the proposed framework showed high accuracy. This means that gets a better prediction performance than other rating predictors.
[23]	2016	Decision Support Systems	Stanford Part-of-Speech (POS) tagger Amazon reviews Phonearena reviews	To propose a method that automatically builds perceptual maps from customer reviews to reduce subjective personal bias.	Author's method called Mining Perceptual Map (MPM) which considers: (1) Weighted LDA (WLDA) (2) Constraint Based LDA (CBLDA)	The method proposed maps in charts properly the different brands considering customer's reviews per feature. Nevertheless, the authors pointed that future research could consider improving the accuracy of this method.
[24]	2016	Electronic Library	Reviews from Amazon Epinions	To propose an Opinion Mining (OM) method based on topic maps to deal with the complexity of natural language	Ontology framework using the following software: (1) ICTCLAS software	The method proposed by the authors improves the accuracy of previous OM.

[25]	2016	International Journal of Systems Science	(not specified)	To generate a portal that contains feature-based products review summary; and develop a predictive model to forecast sales in the following week based on numerical review ratings	(1) Logistic regression (2) Naïve Bayes	The authors obtained an accuracy of 90% in the summarization of the reviews. Nevertheless, the sales forecasting model has not been achieved.
[26]	2016	Engineering Applications of Artificial Intelligence	Pang and Lee subjective dataset MPQA project lexicon dataset Amazon reviews	To present an approach considering product features for a better comprehension.	(1) Co-clustering algorithm	This study has presented a tool addressed to designers to process customer's requirements more efficiently.

Table 7 Literature review from the selected articles (own elaboration)

For further research, the LR considers analysing the item from which the reviews are extracted.

Item/s tested by article					
[1]	Not specified	[10]	(1) Movies; (2) Books; (3) electronics; (4) home appliances; (5) sports gears; (6) pets	[19]	Cameras
[2]	Mobile APPs	[11]	Smartphones	[20]	Not specified
[3]	Smartphones	[12]	Smartphones	[21]	(1) Cameras; (2) Smartphones; (3) tablets
[4]	Movies	[13]	(1) Videogames; (2) Books; (3) electronics; (4) Sports; (5) Baby	[22]	(1) Wrist watches; (2) jewellery; (3) Software
[5]	Not specified	[14]	E-reader	[23]	Smartphones
[6]	Toys	[15]	Not specified	[24]	Washing machines
[7]	Cars	[16]	(1) Electronics; (2) movies; (3) hotels; (4) twitter	[25]	Smartphones
[8]	Movies	[17]	Computers	[26]	Smartphones
[9]	(1) Action Figures & Statues; (2) Camera & Photo; (3) Cell Phones; (4) Kitchen & Dining; (5) Skin Care	[18]	(1) Smartphones; (2) movies		

Table 8 Items tested by article (own elaboration)

Conclusions

Most of the articles in table 7 propose models and methodologies to optimize and harness the potential of Sentiment analysis. In addition, the articles propose analysis, approaches and algorithms for improving what has been already proposed in previous literature.

In the LR conducted, it has been seen how ML can be used in e-commerce platforms as a solution for (1) classification, (2) product ranking, (3) prediction, and (4) feature evaluation.

Regarding the techniques used, in the literature it can be found several types. In most of the analysed articles, more than one technique has been performed to compare the effectivity of the method.

Most of the articles have included more than one dataset Some datasets are used for training and testing and others to perform the model. It has been seen frequently in the LR the datasets created by Bing Liu's and NTLK, among others; these two specifically are open and used to train and test the models. To complete the model, it has been used the customers reviews gathered from Amazon and other e-commerce platforms; these have been obtained through web scrapping techniques or platforms that proved scrapping services.

In all the articles analysed, the results showed how the performed technique works and outperforms previous studies.

As seen in table 8, the range of items in which customer reviews have been extracted is varied. Although, the electronics category has been the most explored; this category includes smartphones, cameras, computers, among others.

Regarding the research questions, they can be answered with the results obtained:

Q1: None of the articles from table 5 belongs to a marketing journal. They come from different journals mostly related with Information Management Systems or Computer Engineering.

Q2: The mainstream trend in Opinion Mining research is the combination between different techniques in which the aim is to offer a better result than previous related studies.

Regarding the most reliable techniques, it is not possible to give a concrete answer. It all depends on the characteristics of the dataset and the objective of the study. Nevertheless, the most recent studies have included Deep Learning techniques for dealing with Sentiment Analysis such as Convolutional Neural Network (CNN) or Long-Short Term Memory (LSTM).

Q3: It has been observed how ML is used for Sentiment classification management, product ranking, and for prediction purposes. Nevertheless, it has not been detected prescriptive purposes in any of the selected papers in the LR.

Work completion Schedule

In this study, I have described the different considerations and steps, based on academic literature review, that a marketing director should take into account to successfully

conduct a Business Intelligence process considering the advantages and challenges of the Big Data context.

As described, Opinion Mining and Sentiment Analysis are a powerful tool to canalize and take advantage of the vast amount of data that can be obtained from human generated content collected from datasets like Amazon. Considering customer's reviews from Amazon marketplace, the types of analysis that can be performed are:

1. Level of demand by consumers for a specific product by country
    - From client perspective
    - From provider perspective
2. Weight of opinions using Lexicon approach considering fuzzy logic with updated opinionated data, comparing it with the business performance of the company that sells the specific product.
3. Level of demand once the item has been launched to the market and after certain period of time.
4. Perform a cluster analysis to group products with similar features.

The schedule for the completion of the research will consider the 4 steps mentioned in BI process (Data acquisition, Cleansing, Aggregation and Modelling).

First step (1st year) will be selecting the e-commerce category in which it will be performed the BI process. This step will include also the definition of the objectives as well as the reach taking into consideration the types of analysis and techniques mentioned in the study.

During this period, it will also be performed the Scraping techniques to be able to have current data for the analysis in the following steps.

Second step (2nd year)**:** it will cover the part of data cleansing from the data collected in the previous stage. During this step it will also be prepared the ground and select the appropriate platforms for the following step.

Third step (3rd year)**:** during this year it will be completed the integration and aggregation part to start with the modelling process. It will be performed through ML algorithms as well as Lexicon methods and Fuzzy Logic.

Forth step (4th year)**:** this step will cover the tune-up of the ML and Lexicon methods models from the previous phase and present a proven final Sentiment Analysis model for companies to apply to their business intelligence.

Future research

For future research it can be considered proposing a model to process and analyse data from User Reviews, specifically from Amazon. With an updated and proper model, companies would have a tool that allows them to agilely (1) classify automatically opinions as positive, neutral or negative through Machine Learning and (2) uncover which features are the most valued and which have pushed the consumer to buy.

In addition, considering the current trends in the field of Artificial Intelligence and Machine Learning in the area of analytics, it would be interesting to consider the topics (1) Streaming analytics and (2) Autonomous analytics (Davenport, 2017).

Bibliography

Al-Shabi, M. A. (2020). Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. *International Journal of Computer Science and Network Security*, *20* (1), 51–57.

Al Nuaimi, K., Mohamed, N., Al Nuaimi, M., & Al-Jaroodi, J. (2015). ssCloud: A Smart Storage for Distributed DaaS on the Cloud. In *2015 IEEE 8th International Conference on Cloud Computing* (pp. 1049–1052).

Alavi, M., & Leidner, D. E. (2001). Knowledge Management and Knowledge Management Systems: Conceptual foundations and research issues. *Mis Quarterly*, *25* (1), 107–136.

Ali, F., Kwak, K.-S., & Kim, Y.-G. (2016). Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification. *Applied Soft Computing*, *47*, 235–250.

Appelbaum, D., Kogan, A., Vasarhelyi, M., & Yan, Z. (2017). Impact of business analytics and enterprise systems on managerial accounting. *International Journal of Accounting Information Systems*, *25*, 29–44.

Assante, D., Castro, M., Hamburg, I., & Martin, S. (2016). The Use of Cloud Computing in SMEs. *Procedia Computer Science*, *83*, 1207–1212.

Bryant, R. E., Katz, R. H., & Lazowska, E. D. (2008). Big-data computing: creating revolutionary breakthroughs in commerce, science, and society computing. *Computing Research Initiatives for the 21st Century. Computing Research*

Association.

Calegari, S., & Ciucci, D. (2006). Towards a fuzzy ontology definition and a fuzzy extension of an ontology editor. In *International Conference on Enterprise Information Systems* (pp. 147–158).

Candel Haug, K., Kretschmer, T., & Strobel, T. (2016). Cloud adaptiveness within industry sectors - Measurement and observations. *Telecommunications Policy*, *40* (4), 291–306.

Cox, M., & Ellsworth, D. (1997, October). Application-controlled demand paging for out-of-core visualization. In Proceedings. *Visualization'97 (Cat. No. 97CB36155)* (pp. 235-244). IEEE.

Curuksu, J. D. (2019). *DATA DRIVEN: An Introduction to Management Consulting in the 21st Century*. SPRINGER.

Davenport, T. H. (2017). How analytics has changed in the last 10 years (and how it's stayed the same). *Harvard Business Review*, *28* (08), 2017.

de Pablos, C., López-Hermoso, J. J., Martin-Romo, S., & Medina, S. (2019). *Organización y transformación de los sistemas de información en la empresa*. Madrid: ESIC.

Doan, A., Halevy, A., & Ives, Z. (2012). *Principles of data integration*. Elsevier.

Dubovikov, K. (2019). *Managing Data Science*. Packt Publishing Ltd.

Earley, S., & Bernoff, J. (2020). Is Your Data Infrastructure Ready for AI? *Harvard Business Review*, 1–5. Retrieved from https://hbr.org/2020/04/is-your-data-infrastructure-ready-for-ai

Eberius, J., Thiele, M., & Lehner, W. (2017). Exploratory Ad-Hoc Analytics for Big Data. In *Handbook of Big Data Technologies* (pp. 365–407). Springer.

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research*, *69* (2), 897–904.

Ertel, W. (2017). Machine learning and data mining. In *Introduction to Artificial Intelligence* (pp. 175–243). Springer.

Fernández, A., del Río, S., López, V., Bawakid, A., del Jesus, M. J., Benítez, J. M., & Herrera, F. (2014). Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks. *Wiley Interdisciplinary*

Reviews: Data Mining and Knowledge Discovery, 4 (5), 380-409.

File storage, block storage, or object storage? (n.d.). Retrieved June 7, 2020, from https://www.redhat.com/en/topics/data-storage/file-block-object-storage

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

Gil-Lafuente, A., Garcia-Rondón, I., Souto-Anido, L., Blanco-Campins, B. E., Ortíz-Torres, M., & Zamora-Molina, T. (2017). *La gestión y toma de decisiones en el sistema empresarial cubano*. Barcelona: Real Academia de Ciencias Económicas y Financieras.

Google. (2020). Google Trends. Retrieved May 24, 2020, from https://trends.google.com/trends/explore?date=all&q=big data

Hemmatian, F., & Sohrabi, M. K. (2017). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, 52, 1495-1545.

Henderson, R. ;Clark, K. (1990). Architectural Innovation : The Reconfiguration of Existing Product Technologies and the Failure of Established Firms. *Administrative Science Quarterly, Special Issue*, *35* (1), 9–30.

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., & Guru, S. (2016). *The Age Of Analytics: Competing In A Data-Driven World*. Retrieved from https://www.mckinsey.com/~/media/mckinsey/business functions/mckinsey analytics/our insights/the age of analytics competing in a data driven world/mgi-the-age-of-analytics-full-report.ashx

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. In *Ldv Forum,* 20 (1), 19–62.

Jelonek, D. (2017). Big Data Analytics in the Management of Business. In *MATEC Web of Conferences*, 125(4021), 1-6.

Karacapilidis, N., Tzagarakis, M., & Christodoulou, S. (2013). On a meaningful exploitation of machine and human reasoning to tackle data-intensive decision making. *Intelligent Decision Technologies*, *7* (3), 225–236.

Kaufmann, A., & Gil-Aluja, J. (1986). *Introducción de la teoría de los subconjuntos borrosos a la gestión de las empresas*. Pontevedra: Milladoiro.

Kraska, T. (2013). Finding the needle in the big data systems haystack. *IEEE internet Computing, 17 (1)*, 84-86.

Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, *21* (1), 1–24.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, *6* (70), 1–4.

Lyman, P. (2003). How much information? Retrieved June 21, 2020, from http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

Lynn, T., Rosati, P., Lejeune, A., & Emeakaroha, V. (2017). A preliminary review of enterprise serverless cloud computing (function-as-a-service) platforms. In *2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)* (pp. 162–169).

Marklogic Coorporation. (2019). Training the classifier. *MarkLogic Server -Search Developer's Guide*. Retrieved May 29, 2020, from https://docs.marklogic.com/guide/search-dev.pdf

McCartney, B., & Potts, C. (2020). Natural Language Understanding. In *Supervised Sentiment Analysis* (pp. 1–82). Stanford University. Retrieved from https://web.stanford.edu/class/cs224u/materials/cs224u-2020-sentiment-handout.pdf

Melgrati, I. (2018). Cloud services delivery models. Which can help your business? Retrieved May 28, 2020, from https://imelgrat.me/cloud/cloud-services-models-help-business/

Mohanty, S., Jagadeesh, M., & Srivatsa, H. (2013). *Big data imperatives: Enterprise 'Big Data'warehouse,'BI' implementations and analytics*. Apress.

Mukkamala, R. R., Hussain, A., & Vatrapu, R. (2014). Fuzzy-set based sentiment analysis of big social data. In *2014 IEEE 18th International Enterprise Distributed Object Computing Conference* (pp. 71–80).

Oliveira, T., Thomas, M., & Espadanal, M. (2014). Assessing the determinants of cloud computing adoption: An analysis of the manufacturing and services sectors. *Information and Management*, *51* (5), 497–510.

Patterson, J., & Gibson, A. (2017). *Deep learning: A practitioner's approach*. Sebastopol, EUA: " O'Reilly Media, Inc."

Powell, T. C., & Dent-Micallef, A. (1997). Information technology as competitive advantage: The role of human, business, and technology resources. *Strategic Management Journal*, *18* (5), 375–405.

Prathi, J. K., Raparthi, P. K., & Gopalachari, M. V. (2020). Real-Time Aspect-Based Sentiment Analysis on Consumer Reviews. In *Data Engineering and Communication Technology* (pp. 801–810). Springer.

Press, G. (2014). 12 big data definitions. Retrieved May 16, 2020, from https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#4afd861e13ae

Ramzan, B., Bajwa, I. S., Jamil, N., Amin, R. U., Ramzan, S., Mirza, F., & Sarwar, N. (2019). An Intelligent Data Analysis for Recommendation Systems Using Machine Learning. *Scientific Programming*, *2019*.

Raschka, S. (2015). *Python machine learning*. Packt Publishing Ltd.

Sánchez, P. R. P., & Correia, M. B. (2016). The Paradigm of the Cloud and Web Accessibility and its Consequences in Europe. In *Proceedings of the 7th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion - DSAI 2016* (pp. 362–369). New York, New York, USA: ACM Press.

Scott, G. (2020). Artificial Intelligence. Retrieved June 2, 2020, from https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp

Serrato, M., & Ramirez, J. (2017). The strategic business value of big data. In *Big Data Management* (pp. 47–70). Springer.

Shafaee, A., Issa, H., Agne, S., Baumann, S., & Dengel, A. (2014). Aspect-based sentiment analysis of amazon reviews for fitness tracking devices. In *In Pacific-Asia Conference on Knowledge Discovery and Data Mining.* (Vol. 8643, pp. 50–61). Kaiserslautern, Germany: Springer, Cham.

Shayaa, S., Jaafar, N. I., Bahri, S., Sulaiman, A., Wai, P. S., Chung, Y. W., Al-Garadi, M. A. (2018). Sentiment analysis of big data: Methods, applications, and open challenges. *IEEE Access*, *6*, 37807–37827.

Sheela, L. J. (2016). A review of sentiment analysis in twitter data using Hadoop. *International Journal of Database Theory and Application*, *9* (1), 77–86.

Shim, J. P., French, A. M., Guo, C., & Jablonski, J. (2015). Big data and analytics:

Issues, solutions, and ROI. *Communications of the Association for Information Systems*, *37* (1), 797–810.

Simalango, M. F., Kang, M.-Y., & Oh, S. (2010). Towards Constraint-based High Performance Cloud System in the Process of Cloud Computing Adoption in an Organization. *ArXiv.Org*, 45–55.

Simpson, J., & Weiner, E. (1989). Oxford english dictionary. Retrieved May 15, 2020, from https://www.oed.com/view/Entry/18833?redirectedFrom=big+data

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286.

Statista. (2018a). Number of installed cloud workloads worldwide, from 2015 to 2021, by service type (in millions). Retrieved June 8, 2020, from https://www-statista-com.sire.ub.edu/statistics/633873/worldwide-cloud-workloads-by-service-type-installed/

Statista. (2018b). Volume of data/information created worldwide from 2010 to 2025 (in zetabytes). Retrieved May 25, 2020, from https://www.statista.com/statistics/871513/worldwide-data-created/

Statista. (2020). How do you search for specific information on a product that you want to buy? Retrieved June 6, 2020, from https://www.statista.com/forecasts/997051/sources-of-information-about-products-in-the-us

Subroto, A., & Apriyana, A. (2019). Cyber risk prediction through social media big data analytics and statistical machine learning. *Journal of Big Data*, *6* (1), 50.

Szwacka-Mokrzycka, J. (2015). an Interdisciplinary Approach To Marketing. *Annals of Marketing Management & Economics*, *1*(1), 85–92.

Tilly, C. (1984). The Old New Social History and the New Old Social History. *Fernand Braudel Center*, *7* (3), 363–406.

Tutorial, T. H. (2020). 5 Top Hadoop Alternatives to Consider in 2020. Retrieved June 2, 2020, from https://www.hdfstutorial.com/blog/top-hadoop-alternatives/

Vivancos, D. (2018). *Big Data: Hacia la inteligencia artificial*. Barcelona: The Valley Digital Business School.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34* (2), 77–84.

Webster, M. (n.d.). Fuzzy logic. Retrieved June 3, 2020, from https://www.merriam-webster.com/dictionary/fuzzy logic

Woods, D. (2011). Big Data Requires a Big, New Architecture. Retrieved June 1, 2020, from https://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/

Zomaya, A. Y., & Sakr, S. (2017). *Handbook of big data technologies*. Cham, Switzerland: Springer.

Appendix

[1] Mowlaei, M. E., Abadeh, M. S., & Keshavarz, H. (2020). Aspect-based sentiment analysis using adaptive aspect-based lexicons. Expert Systems with Applications, 148, 113234.

[2] Malik, H., Shakshuki, E. M., & Yoo, W. S. (2020). Comparing mobile apps by identifying 'Hot'features. Future Generation Computer Systems, 107, 659-669.

[3] Yang, Z., Ouyang, T., Fu, X., & Peng, X. (2020). A decision-making algorithm for online shopping using deep-learning–based opinion pairs mining and q-rung orthopair fuzzy interaction Heronian mean operators. International Journal of Intelligent Systems, 35(5), 783-825.

[4] Ghorbani, M., Bahaghighat, M., Xin, Q., & Özen, F. (2020). ConvLSTMConv network: a deep learning approach for sentiment analysis in cloud computing. Journal of Cloud Computing, 9(1), 1-12.

[5] Hammou, B. A., Lahcen, A. A., & Mouline, S. (2020). Towards a real-time processing framework based on improved distributed recurrent neural network variants with fastText for social big data analytics. Information Processing & Management, 57(1), 102122.

[6] Wang, W. M., Wang, J. W., Li, Z., Tian, Z. G., & Tsui, E. (2019). Multiple affective attribute classification of online customer product reviews: A heuristic deep learning

method for supporting Kansei engineering. Engineering Applications of Artificial Intelligence, 85, 33-45.

[7] Liu, Y., Jiang, C., & Zhao, H. (2019). Assessing product competitive advantages from the perspective of customers by mining user-generated content on social media. Decision Support Systems, 123, 113079.

[8] Patel, N. V., & Chhinkaniwala, H. (2019). Investigating Machine Learning Techniques for User Sentiment Analysis. International Journal of Decision Support System Technology (IJDSST), 11(3), 1-12.

[9] Lee, H. C., Rim, H. C., & Lee, D. G. (2019). Learning to rank products based on online product reviews using a hierarchical deep neural network. Electronic Commerce Research and Applications, 36, 100874.

[10] Kincl, T., Novák, M., & Přibil, J. (2019). Improving sentiment analysis performance on morphologically rich languages: Language and domain independent approach. Computer Speech & Language, 56, 36-51.

[11] Zhang, L., Chu, X., & Xue, D. (2019). Identification of the to-be-improved product features based on online reviews for product redesign. International Journal of Production Research, 57(8), 2464-2479.

[12] Abulaish, M., & Bhardwaj, A. (2019). OMCR: An Opinion-Based Multi-Criteria Ranking Approach. Journal of Intelligent & Fuzzy Systems, 36(1), 397-411.

[13] Zhang, S., & Zhong, H. (2019). Mining users trust from e-commerce reviews based on sentiment similarity analysis. IEEE Access, 7, 13523-13535.

[14] Zhou, F., Jiao, J. R., Yang, X. J., & Lei, B. (2017). Augmenting feature model through customer preference mining by hybrid sentiment analysis. Expert Systems with Applications, 89, 306-317.

[15] Nasim, Z., & Haider, S. (2017). ABSA toolkit: An open source tool for aspect-based sentiment analysis. International Journal on Artificial Intelligence Tools, 26(06), 1750023.

[16] Choi, Y., & Lee, H. (2017). Data properties and the performance of sentiment classification for electronic commerce applications. Information Systems Frontiers, 19(5), 993-1012.

[17] Zou, H., Gong, Z., & Hu, W. (2017). Identifying diverse reviews about products. World Wide Web, 20(2), 351-369.

[18] Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. Expert Systems with Applications, 69, 214-224.

[19] Vinodhini, G., & Chandrasekaran, R. M. (2017). A sampling-based sentiment mining approach for e-commerce applications. Information Processing & Management, 53(1), 223-236.

[20] Farooq, U., Nongaillard, A., Ouzrout, Y., & Qadir, M. A. (2016). A feature-based reputation model for product evaluation. International Journal of Information Technology & Decision Making, 15(06), 1521-1553.

[21] Zhou, Q., Xia, R., & Zhang, C. (2016). Online shopping behavior study based on multi-granularity opinion mining: China versus America. Cognitive Computation, 8(4), 587-602.

[22] Cosma, G., & Acampora, G. (2016). A computational intelligence approach to efficiently predicting review ratings in e-commerce. Applied Soft Computing, 44, 153-162.

[23] Lee, A. J., Yang, F. C., Chen, C. H., Wang, C. S., & Sun, C. Y. (2016). Mining perceptual maps from consumer reviews. Decision Support Systems, 82, 12-25.

[24] Xia, L., Wang, Z., Chen, C., & Zhai, S. (2016). Research on feature-based opinion mining using topic maps. The Electronic Library, 34(3), 435-456

[25] Kangale, A., Kumar, S. K., Naeem, M. A., Williams, M., & Tiwari, M. K. (2016). Mining consumer reviews to generate ratings of different product attributes while producing feature-based review-summary. International Journal of Systems Science, 47(13), 3272-3286.

[26] Jin, J., Ji, P., & Kwong, C. K. (2016). What makes consumers unsatisfied with your products: Review analysis at a fine-grained level. Engineering Applications of Artificial Intelligence, 47, 38-48.