International Journal of Information Management 44 (2019) 1–12

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Towards a big data framework for analyzing social media content

Jose Luis Jimenez-Marquez*, Israel Gonzalez-Carrasco, Jose Luis Lopez-Cuadrado, Belen Ruiz-Mezcua

Department of Computer Science and Engineering, Universidad Carlos III de Madrid, Leganes, Spain

ARTICLE INFO

Keywords:
Big data framework
Machine learning model
Social media analytics
Hospitality
Yelp

ABSTRACT

Modern companies generate value by digitalizing their services and products. Knowing what customers are saying about the firm through reviews in social media content constitutes a key factor to succeed in the big data era. However, social media data analysis is a complex discipline due to the subjectivity in text review and the additional features in raw data. Some frameworks proposed in the existing literature involve many steps that thereby increase their complexity. A two-stage framework to tackle this problem is proposed: the first stage is focused on data preparation and finding an optimal machine learning model for this data; the second stage relies on established layers of big data architectures focused on getting an outcome of data by taking most of the machine learning model of stage one. Thus, a first stage is proposed to analyze big and small datasets in a non-big data environment, whereas the second stage analyzes big datasets by applying the first stage machine learning model of. Then, a study case is presented for the first stage of the framework to analyze reviews of hotel-related businesses. Several machine learning algorithms were trained for two, three and five classes, with the best results being found for binary classification.

1. Introduction

Social media companies became popular with the advent of the Internet in the late 1990s. In those early days, users expressed their feelings about the products they bought or the services they used commonly through blogs, web chats in dedicated forums or via email to the provider. As e-commerce continued evolving, enterprises such as Amazon and the Internet Movie Database (IMDb) included for every item (e.g. CDs, books, DVDs, movies, TV series, etc.) a means for registered users to be able to interact among themselves and to share opinions about their buying experiences.

Since then, these services have evolved in many ways to offer users more sophisticated methods to enrich the review experience. Some of the add-ons that now come along with the review text are: number of stars on a given scale, number of votes that found the review useful, photo of the reviewer, popularity of the reviewer, number of reviews given by the reviewer, images to illustrate or support the argument, kind of services provided (indicated by the customers), overall rating of the service/product provider, etc.

Many of the features mentioned above have been integrated into services by digital companies such as TripAdvisor, Airbnb, Amazon, Yelp, Cabify, Blablacar, Foursquare and Booking.com. These features generate giant volumes of information that are commonly referred to as Big Data (BD): Petabytes and even exabytes of data that are being generated by these type of enterprises (Gandomi & Haider, 2015). Companies of a minor scale not solely dedicated to digital services are also generating big volumes of data that reach terabytes of data on a regular basis. For further information, Yaqoob et al. (2016) present a robust study of the evolution of BD from its conception to its future challenges, aimed at a more comprehensive understanding of the BD scenario.

Companies and institutions across the world are gaining valuable insights into the massive amounts of the information they have by applying tools and techniques of BD. These techniques are commonly known as Big Data Analytics (BDA) and consist of a set of algorithms, advanced statistics and applied analytics. BDA "refers to the techniques utilized to examine and process BD so that hidden underlying patterns are revealed, relationships are identified, and other insights concerning the application context under investigation are exposed" (Iqbal, Doctor, More, Mahmud, & Yousuf, 2016). Modern companies need to have new strategies to handle huge volumes of information and find the hidden knowledge in these data. Several frameworks and methodologies have been proposed to tackle this challenge from scientific and technological perspectives. In Habib, Chang, Batool, and Ying (2016) a BD Reduction

* Corresponding author at: Department of Computer Science and Engineering, Universidad Carlos III de Madrid, Av. de la Universidad, 30, 28911, Leganés, Madrid, Spain.
E-mail address: joseluis.j.marquez@alumnos.uc3m.es (J.L. Jimenez-Marquez).

https://doi.org/10.1016/j.ijinfomgt.2018.09.003
Received 15 May 2018; Received in revised form 13 September 2018; Accepted 13 September 2018
0268-4012/ © 2018 Elsevier Ltd. All rights reserved.

Framework is proposed for decreasing data in the early phases.

To achieve better results in data analysis, complex techniques are being integrated into many analysis models. Lismont, Vanthienen, Baesens, and Lemahieu (2017) conducted a survey to analyze the techniques that "can enhance the decision-making process in companies". In their research they noted that several Machine Learning (ML) techniques are being used for analytics in the organizations, with linear regression and decision trees being the most prevalent. Thus, a first stage of data analysis has been proposed, focused primarily on finding the algorithms that achieve best results for the data available according to the goals of the study.

Social commerce is an area of research with many directions of interest, as stated Lin, Li, and Wang (2017). User Generated Content (UGC) and online reviews are the trends receiving most attention from researchers today. By using BDA and ML, companies can increase their potential advantages and boost their revenues by enhancing relation with clients with customized offers according to their records. Hence, a state-of-the-art model able to manage large volumes of information and find valuable insights in data is proposed. Moreover, modern companies need to have as part of their human assets new profiles capable not only of knowing how to handle data, but to find patterns in the information and know how to transform them into new incomes or competitive advantages; this human asset is known as the data scientist (Costa & Santos, 2017; Larson & Chang, 2016).

In the age of social media, users of products and services now prefer to read other users' reviews before deciding to buy a product. Ahmad and Laroche (2017) analyzed a set of Amazon products to study the differences between positive and negative reviews by applying ML techniques to explain customer behavior. In a similar way, this study applies ML techniques to users' reviews of hotel services as a case study to capture the overall sentiment of a business unit; this enables the CEO to know the current image of the company according to their customers' preferences.

The purpose of this paper is to present a computational framework for the management of BD focusing primarily, but not exclusively, on sets of information containing UGC. The two contributions of this work are: 1) a BD and ML framework designed to process both qualitative (i.e. text valuations) and quantitative (i.e. user ratings) information for the predictive analysis of text data is presented, and 2) through a series of ML and Natural Language Processing (NLP) techniques, the framework classifies user reviews into positive or negative in a subset of the Yelp dataset. The results show that high accuracy was achieved for the binary classifier using Multi-Layer perceptron.

The proposed framework is a two-stage model, and consists of: a first stage, where a set of phases are related to managing and processing social media text data to establish a Machine Learning Model (MLM) that can be used in the next stage. The second stage is comprised of BD architecture and data analysis phases that use the previously developed MLM to get results using BDA as well as other BD techniques. This research presents the elements integrating the framework and the results obtained by applying the methodology in the first stage.

This paper is structured as follows: Section 2 explores how BD and ML are shaping the future of social media domains, with emphasis in the tourism sector; Section 3 presents the proposed framework, exposing the characteristics and methodology involved in its design; Section 4 presents the results for the first stage of the framework; in Section 5, results of the study are discussed, and in Section 6 the conclusions of the research are presented.

2. Background

Social media is today becoming the focus of many research studies, mostly because it reaches most of the world's population; many people have access to mobile devices and are also users of social media services. Social media is a great resource for applying BDA (Tan, Blake, Saleh, & Dustdar, 2013) in order to, for example, gain insights into user preferences, explore daily trending, understand the behavior of users with related affinities or analyze habits in population. Social media has the data necessary to analyze these situations: likes, states, text, images, etc. This section presents the theoretical background of the techniques proposed to analyze this data as well as the opportunities in this area.

2.1. Machine learning in social media tourism

Artificial Intelligence (AI) and ML are literally changing everything. It is expected that the 21st century will witness the explosion of all their potential in every aspect of human life. The tourism industry is not an exception since it also needs AI and ML to enhance its businesses' models.

Early studies were carried out by Law, Rong, Quan, Li, and Andy (2011), Lin and Chen (2012) who worked with Hong Kong-related datasets to apply association rules. They first analyzed the behavior of outbound tourism to find the destinations that Hong Kong travelers most prefer. In a second study, they analyzed how to integrate electronic word of mouth in the tourism sector by applying advanced data mining techniques. They identified the characteristics of sharers and browsers, pointing out the underlying features that induce an internet user to rate and share their experiences of past travels, which could help tourism managers to identify potential customers for strategical decision taking. On the one hand, Law et al. (2011) consider the 2005–2009 annual domestic surveys of Hong Kong outbound tourism. Such surveys are related to travelers only visiting this specific destination. Although the information is abundant and heterogeneous, it is limited to the survey's own considerations and purposes and the responses do not consider qualitative opinions or free expressions. At the same time, this paper only considers a specific algorithm (the targeted positive/negative rule discovery) in the context of contrast mining. Rong, Quan, Law, and Li (2012) consider a domestic tourism survey of outbound pleasure travel in Hong Kong. That survey is related to past experiences when traveling, and their web experience before the travel journey. A part of this survey were open answers to share more information about their reasons for travel. Even though this study is an improvement with respect to the authors' previous research paper, it does not take into consideration the information expressed in free form. The paper is also focused, as is the previous one, on showing the effectiveness of association rules to analyze the information contained in the surveys, and therefore only considers one algorithm.

An important research study by Xiang, Du, Ma, and Fan (2017) revealed that the characteristics of the datasets that have been used in many studies, namely TripAdvisor, Expedia, and Yelp, can vary significantly according to the provider, mostly because those datasets contain substantial differences according to a variety of features such as: the popularity of the platform, the users for each one of these platforms and the size of the hotels that are commonly rated on each of these sites. That research analyzes the information in these three data sources to study the differences between them, concluding that future work should explore relationships between review content and sentiment.

Other studies such as the one by Silva and Zhao (2015) made use of mixed types of data to conduct a study of tourist behavior at local destinations by analyzing the walks they go on. That research is mostly focused on pattern recognition and how to analyze information related to "tourist walks". Although the type of data sources is not of the type analyzed for our framework, it would be very helpful for further studies to integrate information on the places where the tourists have been to get information about the most visited places in a location.

A very interesting study was carried out by Deng and Robert (2018), who used a Flickr's dataset to analyze the information contained in photos of New York City taken by both tourists and local advertisers. The authors found the places most visited by tourists; consequently, these methods could help Destination Marketing Organizations (DMO) to find the most popular places in their cities and then be able to offer

customized marketing ads. In the mentioned study, the researchers used a Naïve-Bayes classifier to analyze the photos; for future work, other classifiers could be integrated into similar studies to enhance and compare previous and new results. That research also relates the images to the users' feelings and categorizes them according to users' sensations. The authors adopted the naïve Bayes classifier in their model. They also suggest that their research could be affected by the sparseness of comments, but do not provide a medium to solve this fact.

The aforementioned studies show some of the research carried out until now in ML for tourism, whose results have provided important contributions to computer sciences as well as to other domains. The framework in this paper proposes to build an MLM able to analyze text data and to reuse this model at a later time by applying most of it in a BD architecture.

2.2. Big data in social media tourism

Tourism is a growing industry; every year millions of tourists around the world visit a tourist destination at least for one night. The 21st century tourist has different habits when compared to a tourist of 20 or 30 years ago. At the very start of choosing a destination, today's tourist looks for tips through reviews in platforms such a TripAdvisor or Yelp. When booking a flight, users can search in engines like Kayak or Skyscanner. During the visit, tourists post pictures and comments in social networks such as Facebook, Twitter or Instagram. After the stay, the tourist continues to write reviews and grade the services used during the trip such as those of the hotel, restaurants and transportation. All these actions create a vast amount of information every day which many times goes unnoticed for CEOs, which could put the company at a disadvantage with its competitors.

As previously stated, this "modern age" amount of information is known as BD, and many companies are taking competitive advantage of it (Kubina, Varmus, & Kubinova, 2015).

As such, modern tourism companies are applying BDA-related techniques (Xu, Wang, Li, & Haghighi, 2017) to get hidden insights into their data or to better know the "likes" and "dislikes" of users. However, they do not make use of ML techniques, but rather use NLP techniques, particularly: latent semantic analysis, singular value decomposition and regression.

Existing studies have found that by taking advantage of BDA infrastructure, core stakeholders can have "real-time knowledge on tourists' on-site behavior at tourism destinations" (Fuchs, Höpken, & Lexhagen, 2014). The authors construct an interesting theoretical background for a "knowledge destination framework" which includes an architecture that describes a series of components for the analysis of their data. Some of these components are similar to the ones proposed by the proposed framework. While the solution presented by these authors is very helpful to gain interesting insights into their data, this study only considers methodology and data that serve the purposes of a specific business intelligence approach.

BDA can also be helpful in the tourism industry to predict the volumes of tourists arriving to a certain location, as established by Li, Pan, Law, and Huang (2017), where the authors propose a methodology to predict these estimates. This study analyzes the words that users query in browsers searching for a specific location. Based on these results, they forecast what the demand for tourist services would be for the coming season. However, they do not consider user reviews. The authors even recognize that this study could be analyzed by ML algorithms as artificial neural networks, and vector autoregressive models. Furthermore, they do not propose a framework on how to integrate their techniques, nor do they consider integrating the analysis in real-time big data tools.

Accordingly, BD and BDA are a set of techniques that, if applied effectively to contribute to the objectives of the company, can produce results able to help the organization obtain competitive improvements or substantial gains. At the core of this BD infrastructure lies the most valuable asset for every company which is information; that is why this framework is focused on the efforts of how to get the most significant insights from data by also having a proper model for management of large volumes of unstructured information.

2.3. Big data frameworks

There is a growing interest about data management frameworks as modern companies are not only interested in collecting digital data and having it stored without further exploitation but are now interested in how they can translate their data into significant results. An example of these frameworks is the one by Vajirakachorn and Chongwatpol (2017) where the authors propose a framework for a local food festival in Thailand. However, their framework does not consider the analysis of textual data since it is intended to work with structured data. The BD framework proposed by Habib, Chang, Batool, and Ying (2016) has as its main purpose the "big data reduction at the customer end". Even though that framework considers more elements regarding the BD infrastructure and methodology, it is particularly focused on data reduction for lowering cost when dealing with BD and cloud computing transactions. In short, the aforementioned BD framework does not consider the integration of ML and NLP for data analysis. The research framework presented by Yuan, Xu, Qian, and Li (2016) considers tourist information crawled from travel blogs. The authors then "implement the frequent pattern mining method" to identify a city's popular locations by creating word vectors to construct a word network. Even though that framework works with tourist information extracted from reviews, it is limited to this domain. In the framework proposed by Chang, Ku, and Chen (2017), the authors present an architecture for aspect-based sentiment analysis. They proved that their convolution tree kernel classification model outperforms other ML methods for sentiment classification, but they do not consider how to integrate these techniques into a BD environment.

2.4. Conclusions of the state of the art

As shown in the previous subsections, despite the fact that there are different approaches in this research area, there are also points still to be covered. On the one hand, AI research from Law et al. and Rong et al. (2011, 2012) is based on specific algorithms. Xiang et al. (2017) provide a review of previous approaches stating that in general that they are restricted to a single source, and propose a set of methodological challenges. One noteworthy challenge is that the review structure and content can be considerably different depending on the platform, and they propose future work in the line of exploring the relationships between review content and sentiment. Silva and Zhao (2015) deal with classification in the tourism domain, but their approach is not based on text and reviews but rather on walking information. Deng and Robert (2018) take into account the comments on the selection of photos but their study is limited to the photo domain.

On the other hand, Big data approaches like Xu et al. (2017) apply NLP techniques for identification of key attributes, but do not include ML approaches or a framework to be applied. Fuchs et al. (2014) propose a big data knowledge infrastructure but restricted to the tourism domain. Finally, Li et al. (2017) provide tourism demand forecasting based on search trends but do not take into account user reviews. With respect other frameworks in the state of the art, they cover several areas and concrete domains but do not provide an integral approach. To the best of our knowledge, a framework for integrating natural language processing, machine learning and big data techniques for social media content in a methodological way, and not restricted to specific algorithms or domains, has not been found in the literature review.

3. Data integration and big data analytics framework

This section presents the conceptual aspects of the proposed

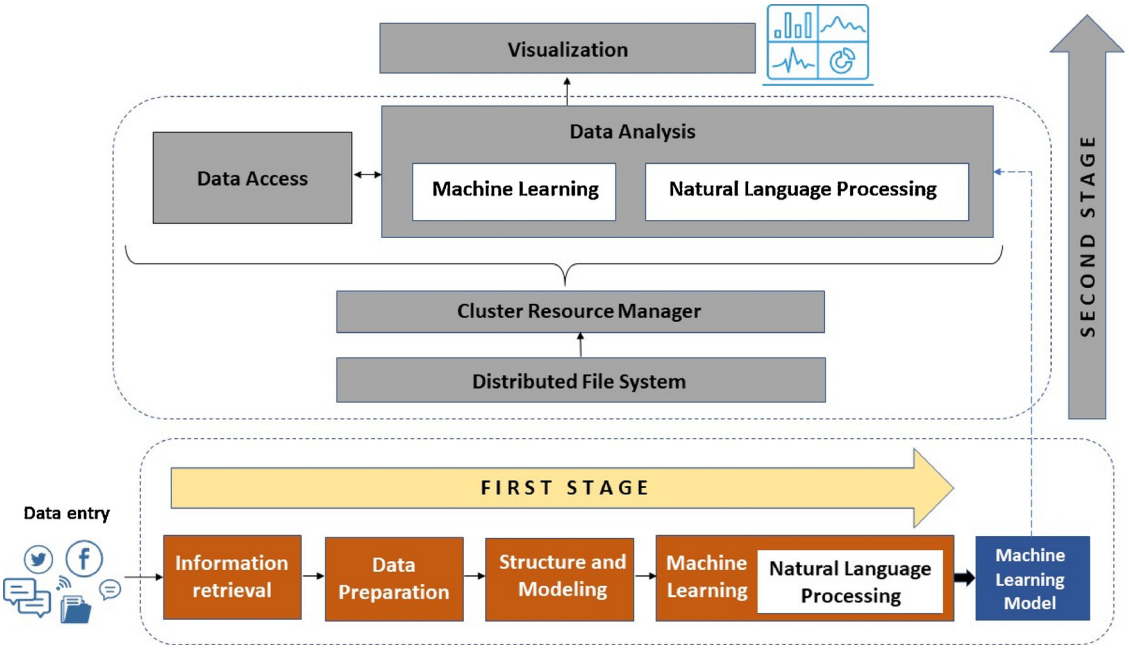

Fig. 1. Data integration and Big Data analytics framework.

framework. The framework is composed of two stages: a data integration stage and a BDA stage. The framework is shown in Fig. 1.

BD is gaining importance, and so companies can leverage on available data to exploit new opportunities and gain important knowledge of the insights hidden in values (Raguseo, 2018). The management of large volumes of information and subsequent analysis is a major issue that must be taken into consideration, whose benefits can be reflected in getting an edge on competitors or in innovations in the organization's inner processes. Considering the two stages of the framework, it can be summed up that it takes as input large sets of volumes of unstructured textual information, and then applies an MLM to get meaningful insights, visualizing these results by high-level techniques.

3.1. First stage

The first stage of the framework proposes a series of phases that represent the sequencing of data, from the data entry to building a MLM. The purpose of this stage is to analyze data that could be either Big or Small Data (BSD); if BD is to be analyzed, a subset of it should be extracted from the BD sources for better results. The advantage of this approach is that, in an intermediate stage, a conception of data structure and content can be obtained. Moreover, this framework proposes the analysis of text data; other types of data sources could be better analyzed by applying different models.

3.1.1. Phase 1: data entry

The data entry to the framework comes from text data sources: unstructured data. These data can be data from social media sources, blogs, chats or micro blogging services. In this paper, as a study case, the data to be analyzed is tourism data, specifically users' reviews on hotels and resorts, but the framework would work well with any data source having social text data.

3.1.2. Phase 2: information retrieval phase

The initial phase of the framework is the information retrieval, which involves carrying out the necessary measures to collect relevant data, even though many companies will use their own text data sources. Companies amass information to develop broad research studies i.e. from customer reviews or sensor data. The information retrieval process must ensure that collecting irrelevant data is avoided. The proper strategies for data collection help to reduce costs by easing the computational and storage burden of data centers (Habib et al., 2016). Data collection techniques contribute to obtaining diverse information to feed BD warehouses. Such techniques are built in-house (i.e., web scrapers) or acquired by a third party. According to Olmedilla, Martínez-Torres, and Toral (2016) data collection can be done using the Application Programming Interfaces (APIs) provided by social networking sites as YouTube, Flickr or Amazon. These authors propose a relevant methodology aimed at gathering data directly from websites with UGC.

Data collection activities must ensure that all the relevant information to conduct the study has been gathered, however, noisy data or lack of data may occur. To avoid data shortage, one possible measure is to apply engineering techniques to collect the data. A return to this phase is always feasible from latter phases, but this would imply an impact on the project's timeline, therefore delaying its completion. This phase's output directly feeds the data preparation phase, since it is responsible for providing the data that will be used in the following phases.

3.1.3. Phase 3: data preparation phase

The activities carried out in the information retrieval phase were related to BSD gathering taken from various sources: At this point, these sets of information must be cleaned and prepared for further analysis.

The second phase (data preparation phase) is the most important in a data project; other associated phases are data preprocessing and integration operations, as stated by Habib et al. (2016). A great part of the time that data science teams invest in is the preparation of data. It is estimated that 70% of the total project time is invested in data preparation and 30% in analysis, which can also be referred to as the data "cleaning".

An activity associated to the data preparation phase is the detection of outliers: values that are not uniform with respect to the rest of data. For example, when working with numerical values these should normally be within a range that satisfies the conditions of the study. However, values that exceed well above or far below the expected indicators are commonly found. Also, when dealing with text values or strings, many of these are null or correspond to different languages, with the consequent change of the characters. In any of these cases as in

many others, the data preparation must include the use of various techniques (mostly computational) to give special treatment to those records. Failing to detect and deal with abnormal data will lead to a significant deviation in the study results.

The output of the data preparation phase is the information in a state of having a homogeneous structure and a series of values in data that suit the objectives of the study as well as not having outliers in present data. This phase will allow the information to be read by BSD techniques, facilitating this process by not finding noise in data. The data preparation phase could be summarized as follows: the original information is "cleaned up" and used in the next phase, responsible for analyzing its structure to create a schematic model that represents the whole set of information.

3.1.4. Phase 4: structure and modeling phase

The purpose of the structure and modelling phase is to obtain the structure of these volumes of information by establishing a data model that defines their inner relations and content. If data are single-sourced, a data model will be easy to create by inspecting all the elements, querying the data and analyzing their relations. Otherwise, random queries or direct data selection will have to be executed on the data objects to identify: the elements that constitute the information, the relations between them and the possible types of data.

The structure and modelling phase has many similarities to the process of designing relational databases, where a structure of related tables and their fields is the final product. The process of structuring and modeling BSD becomes more complex when the information comes from two or more data sources. This is part of the challenging tasks of the data science team and where most valuable insights are hidden. An approach for facing this challenge is to repeat the process of this phase for every data source. Having all the data models, the external relations that connect each source and their items will be established.

When the data model is created some considerations must be taken into account since this phase implies the process of creating a structured model out of unstructured data. In relational databases, data is modeled into a set of tables containing rows and columns, which could be a good fit for some unstructured data sources, even though data coming from social networks such as Facebook or Twitter are oriented to a graph model. Thus, a good design for modelling unstructured data should consider including distinct models, but mainly relational and graph models. The output of this phase is to gain knowledge about data: their nature, their inner structure, the set of values for each field and the possible relations between data objects. This data model will serve as a base to apply techniques of analysis that allow a deeper understanding of data.

3.1.5. Phase 5: machine learning and natural language processing phase

Thais phase proposes that ML joined with NLP techniques will work better for analysis of the unstructured text data obtained at the data preparation phase. ML is revolutionizing every aspect in many areas of science, from image recognition to autonomous vehicles. NLP is an area of computer science that analyzes natural language, helping humans to communicate with computers as they do with other humans. Along these lines, data analysis can perform better when applied to data by using ML and NLP techniques.

Although text data can be analyzed by other techniques (Colace, De Santo, Greco, Moscato, & Picariello, 2015; Li et al., 2017), ML and NLP techniques to the existing information is proposed to be applied for the following reasons: first, because today these are the most accepted and proven techniques by computer science researchers, and secondly, because of the existence of a scientific and technological community that supports and endorses these technological areas (Etaiwi & Naymat, 2017; Ravi & Ravi, 2015; Xiang et al., 2017).

NLP is also considered in this phase since a good deal of data analysis research is being carried out by applying diverse ML algorithms. However, these also rely on many computational packages developed specifically to perform natural language tasks. Accordingly, when NLP is integrated into an ML algorithm, the best results in prediction could be achieved. There are many applications for ML and NLP data analysis. Some of the most relevant ones are: sentiment analysis, recommender systems and user reviews analysis (Alahmadi & Zeng, 2015; Araque, Corcuera-Platas, Sánchez-Rada, & Iglesias, 2017; Chen, Yan, & Wang, 2017). The output of this phase is a method that integrates ML algorithms and NLP packages to perform data analysis.

3.1.6. Phase 6: machine learning model

The last phase of this stage outputs a method that combines ML and NLP techniques focused on text data analysis, such method will constitute the basis of a MLM. This model has the following considerations:

i) The model can be applied to perform data analysis on the existing data to get insights of it, e. g: user likes and dislikes, feature extraction or sentiment analysis. The model's accuracy in predicting results could vary according to the ML and NLP techniques used in the model, however, the tuning of ML algorithms and the integration of many other as math, statistics or computational techniques, is an ever-growing field in text analytics.
ii) The model is intended to be executed on a non-BD architecture, meaning that small data could be analyzed without having high performance resources to perform this task. This also has the advantage that reliable results can be obtained with greater agility by eliminating the barrier of having a very sophisticated computing infrastructure.
iii) This model will be the basis to the ML phase in the second stage since this involves BD Analysis. Therefore, the MLM could be fully or partly applied to the analysis of BD.

3.2. Second stage

As stated previously, BD is a very current trend in science, technology and innovation. Through the analysis of large volumes of data is possible to extract key information that otherwise would not have been obtained by normal means. Leading companies in IT such as Google, Amazon or Facebook are examples where, due to the teamwork of researchers dedicated to BD analysis, strategic results can be achieved that can be a determining factor in establishing a competitive advantage (Özköse, Sertac, & Gencer, 2015). Thus, as shown in Fig. 1, for the second stage of the framework a series of layers aimed at creating a BD Architecture able to perform BD analysis of unstructured heterogeneous data is proposed.

3.2.1. Layer 1. Distributed file system layer

The Distributed File System (DFS) layer is involved in defining a BD architecture. In Oussous, Benjelloun, Ait, & Belfkih, 2017) the authors propose a series of BD products and services. Upon examining the proposal, it was determined that one of these technologies could be integrated into this layer. Spark framework is proposed as the core of the layer, due to the fact that it will provide the next phase with the necessary services to perform data analysis in upper layers. Spark is a BD processing framework that has been used in many research studies not only for text data analysis, but for many other scientific and business applications (Carcillo et al., 2018; Etaiwi, Biltawi, & Naymat, 2017).

It should also be noted that cloud computing solutions (Bayramusta & Nasir, 2016; Lin & Chen, 2012; Sultan, 2013) could be part of this layer. Amazon Web Services and Google Cloud Dataproc support a large set of Spark services, however their pay-as-you-go approach is somewhat prohibitive for small companies. In their deep study of the frameworks, methods and research directions in cloud computing research, Senyo, Addae, and Boateng (2018) present a meta-analysis of 285 articles regarding these topics, and how they relate to technology or business domains, showing how cloud computing is being utilized in

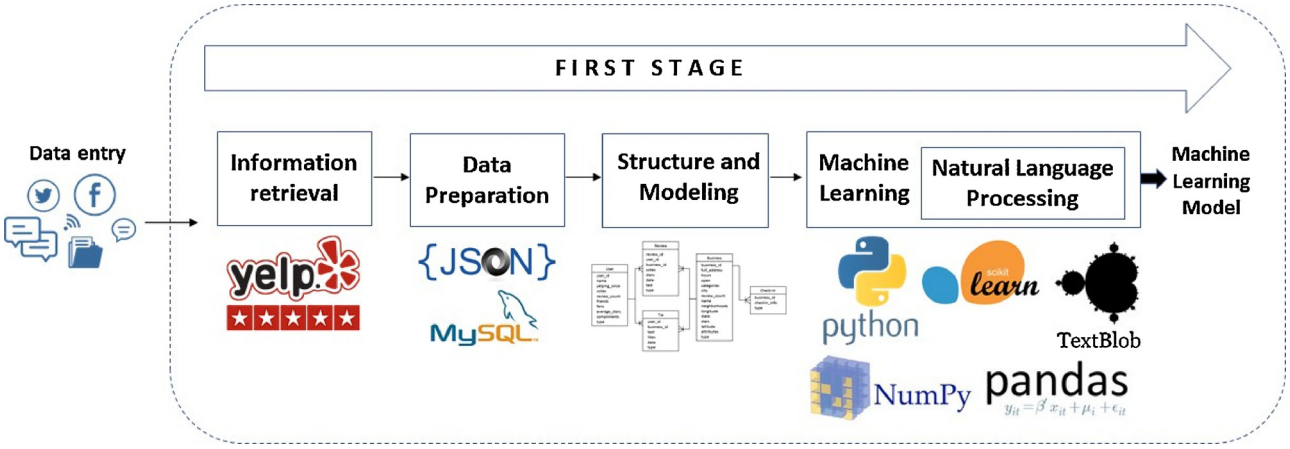

Fig. 2. First Stage phases in detail.

these fields and the challenges their professionals are facing.

3.2.2. Layer 2: cluster resource manager layer

If the DFS is implemented locally, a Cluster Resource Manager (CRM) is required to manage the execution of the tasks distributed among the nodes. A cluster is a collection of computers connected in a network working together to execute several tasks in parallel (Dar, 2016). The cluster allows many computers to be connected, using all their processing power to solve a problem, as well as ensuring fault tolerance since nodes with malfunctions can easily be replaced by others. Therefore, it is suggested that a cluster be used to increase computing power, since BD techniques are created to perform parallel and distributed computing.

A cluster of computers is organized by a central node called the resource manager, which is an activities orchestrator that executes jobs and monitors node performance and operations. If the cluster is to be implemented in the local computing infrastructure, Apache Mesos is proposed as the CRM (Reuther et al., 2018). In the cited paper. the authors compared several CRM, finding Mesos to have the best overall performance. While setting up a local cluster can take considerable time and effort, cloud computing offers many advantages since available services set up clusters in a quick and efficient manner.

3.2.3. Layer 3. Data access layer

Since the DFS and CRM layers have been defined, the data access layer will determine how to manage the access to data being loaded to nodes at a later time. Regarding the inner structure of data, this is managed by the BD system, also supporting a set of languages to access data, such as Python, Java, R or NoSQL, among others. The output of this layer are the data in the DFS's internal format, which will be analyzed by BD techniques in the next layer.

3.2.4. Layer 4. Data analysis layer

The three previous layers were involved in defining a BD architecture, capable of only accessing the existing data, but not of performing any analysis. For this reason, the framework includes the data analysis layer. As this layer is where most of the BD analysis is being conducted, thus the research proposes to apply the ML techniques available in many BD systems (in the case of Spark, MLlib is the ML library) to perform such analysis. One of the framework's proposal was to conduct data analysis of large volumes of text information in a non-BD environment to create an MLM. This model is now expected to be useful in this layer by integrating it into a BD analysis task having a different ML library.

This approach, however, has a limitation: special care must be taken when selecting the ML algorithms to be used (Shirdastian, Laroche, & Richard, 2017), since the ML algorithms in the selected BD system might not support the algorithms in the MLM of the first stage. Since the BD technology is constantly growing, many ML algorithms for BD-like execution are still in the development stage. To tackle this problem, each node can use a single-node learning library (e.g., Weka or SciKit-Learn). Moreover, inside this layer there is a NLP module to perform further analysis of text data when required.

3.2.5. Layer 5. Visualization layer

The previous layer produces as output a series of values representing a behavior that is being measured, which is usually defined at the beginning of the study. To represent these values, it will be necessary to use visualization techniques that allow data analysis on a large scale in a more comprehensible way. Data visualization is a series of techniques for better understanding of information and extracting knowledge from it. It is also used as a presentation tool for the purposes of illustration, explanation and communication of results (Chang et al., 2017; Cybulski, Keller, Nguyen, & Saundage, 2015).

Visual analytic techniques provide stakeholders with powerful elements for decision making since this layer is involved with crafting state-of-the-art charts that allow quick understanding of huge volumes of information in an intuitive and flexible manner by visualizing and interacting with data. The output of this layer is a series of methods developed for the dynamic visualization of large volumes of text information.

4. Results

This paper is part of a research project whose goal is to analyze users' reviews of tourism services, particularly those of hotel and resort services. The interest lies in answers to the following questions: 1) What is the polarity of the sentiment for each group of reviews grouped by stars?; 2) What are the main words that customers express in their reviews? Accordingly, the research aims to make use of the proposed framework to analyze BSD by applying this methodology. This section shows the results obtained following the steps (phases) detailed in the first stage. A panorama of the source material and packages utilized in this stage is shown in Fig. 2, which is explained below.

4.1. Data and methods

Most companies deny the access to their data by external researches due to privacy issues. Other state of the art approaches, try to take the public available information by means of Web Scrapping techniques (Li, Ott, & Varadarajan, 2013). However, the rights to process the data obtained by these means from social web sites are not clear. Yelp corporation publishes a dataset for academic purposes. Yelp is a multinational that develops Yelp.com and the Yelp app, which publishes

reviews about local businesses as well as the Yelp online reservations services. The Yelp dataset has been used in many research studies (Ngo-Ye & Sinha, 2014; Pranata & Susilo, 2016; Xu & Yin, 2015), and therefore it is generally accepted by the computer science community.

To explore the content of the Yelp dataset (Yelp, 2016), it was necessary to build a script to read and fix the contained files, since they did not comply with the JSON definition. After having read the content of the files, it was determined to export these to a relational database for better data analysis.

ML can be done in languages like Python or R. Because of prior experience, it was decided to use Python as the programming language, but nevertheless the use of R as a second platform for ML programming is encouraged. Packages such as NLTK (Natural Language Toolkit) and Pandas (data manipulation and analysis) were used for the NLP task and data preprocessing. Finally, Precision & Recall analysis will be applied in order to analyze the results obtained.

4.2. Information retrieval

At this stage, data can be collected in different ways. The main goal of this step is to take the data and make it available for further processing. Building a web crawler is a common practice to collect data (Guo, Barnes, & Jia, 2017; Hu & Chen, 2016; Marrese-Taylor, Velásquez, Bravo-Marquez, & Matsuo, 2013). A crawler can get public data available and store it in order to be processed in the next steps. Other way is using datasets published for academic purposes, as previously mentioned. Both approaches are supported by the framework and the way the data is obtained is not relevant to the framework. Thus, Yelp dataset was chosen to conduct the study. Yelp dataset is greater than 2GB being one of the bigger ones available for research purposes, and it also contains natural language comments related to a 1–5 scale valuation. These facts make the dataset adequate for testing the proposed framework.

4.3. Data preparation

In this phase, the content of the JSON files of the Yelp dataset was analyzed, which consists of five files: Review, Business, User, Tip and Check-in. As mentioned before, these files were exported to a database. However, this step is merely optional since many organizations can find that due to their day-to-day operations, it is neither possible nor practical to perform this operation. This is not a drawback since the JSON data reading techniques can easily do this task; it was done this way for better data analysis. The file processing has been implemented by means of Python scripts: these processes fix format issues in the JSON files and they also amend lost values or other anomalies (Habib et al., 2016). Finally, these processes prepare the data to be processed in the next phase.

4.4. Structure and modeling

The structure and modeling phase was comprised of analyzing the existing tables. Then the relations shown in the Entity-Relation (E–R) model of Fig. 3 were discovered:

As can be seen in Fig. 3, several relations between the entities of the model could be established. With this information in hand, a substantial understanding of the data was achieved. For example, it was found that the review and tip tables are more related to other tables in the E–R model, and a review registry does not point directly to a certain category, but rather is related to a business unit which in turn is related to many categories through the *categories* field. The categories field is user-described, meaning that users indicate what the business' services are. Therefore, a hotel can be related along with other categories, for example: casinos, tour, food, bar, event, wine, etc. A relational database was created in MySQL according to this structure in order to process the dataset as well as to refine and retrieve the data related to each record (text reviews, number of stars, etc.).

4.5. Machine learning and natural language processing

As previously stated, our goal was to study reviews related to tourism services, specifically hotels. Thus, to distinguish which reviews were related to hotels, a subsample of reviews was taken by relating this table to business in SQL language. Then, from this subsample, only the "stars" and "text" fields were considered for the ML phase. These fields are of major importance since they reflect how the user rates services in both quantitative and qualitative aspects. For a same business unit, users express their feelings in different ways, making it of great interest to study how words and expressions used are related to the "stars" assigned.

As noted, for the ML phase it was necessary to have the information of reviews (stars, text) in a file. However, this data needed additional cleaning since there were many elements encountered in web forms (special characters, blanks, nulls). This data cleansing was done by executing a SQL instruction to eliminate noisy data. Stop words and punctuation were removed from the reviews, which were also transformed to lower-case.

At this phase, PLN techniques are incorporated into the framework to preprocess the available corpus for allowing more adequate processing by means of ML techniques. PLN is related to disciplines such as sentiment analysis where both PLN and ML techniques come together. Therefore, PLN allows to complete this stage of the framework giving a greater sense to the subsequent tasks of ML. Python and the NLTK packages have been applied for developing this stage. In the construction of the ML model, Python or R were options to consider. The final ML development has been based on Python in order to maintain the previous processes in the same language. NLTK and Pandas packages (for data manipulation and analysis), as well as the Scikit-learn tool, an open source Project widely used by industry and academy (Mueller & Guido, 2016), have been applied for the development of the ML processes.

After running preliminary models on the data, it was discovered that the distribution of samples was biased towards reviews having one or four-to-five stars, making those with two and three stars the least present in samples. Such biasing was causing the MLM to focus more on the edges and less in the center, and so it was decided to choose a balanced sample of reviews. The final dataset consisted of 66,410 reviews, with 13,282 reviews for every star; 75% of the total was for training and 25% for testing phases.

4.6. Machine learning model

The first aspect to consider in the model was what vectorizer (which transforms words into vectors) to apply to the data. Having CountVectorizer (Pedregosa et al., 2011; Scikit-learn CountVectorizer, 2018) and TfIdfVectorizer (Pedregosa et al., 2011; Scikit-Learn TFIDFVectorizer et al., 2018) as the main choices, both were tested, finding that the latter performed better for data. The next step was how to set up the TfIdfVectorizer: *i)* data was stemmed via the tokenizer to use text in its root form, *ii)* terms that appear in more than 50% of the documents or in less than five documents are ignored, *iii)* the maximum number of features to be considered are 30,000, and *iv)* n-grams were considered in the range from 1 to 2.

There were two reasons to consider 30,000 as the maximum number of features: first, the raw number of features was above 100,000, leading to low performance and prediction of the ML classifier; second, because it was the most optimal value that fit the data. These considerations in the vectorizer and prior data preprocessing help to avoid overfitting since many tests were executed with different classifiers and combinations of data (as later explained) and it was not necessary to tune a new vectorizer for every test.

The parameters selected for TfIdfVectorizer try to avoid overfitting.

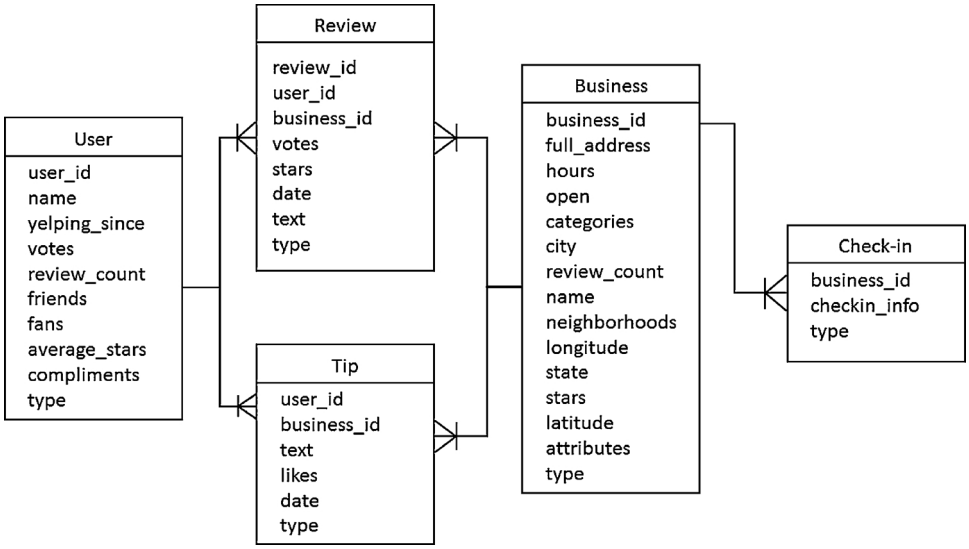

Fig. 3. Yelp Entity-Relation model.

The Python tokenize method is based on the SnowballStemmer from NLTK library, which lowercases each string and translate each word to its root form: it contributes to decrease the corpus size (features), focusing on find the most common terms in the text.

The classifiers used for the test were: Multi-layer Perceptron (MLP), Support Vector Classifier (SVC), Logistic Regression (LR), Linear Support Vector Classification – L1 and L2 penalties (LSVC L1 & L2), Linear classifier with Stochastic Gradient Descent training – L1 and L2 penalties (SGD L1 & L2) and Naive Bayes (NB). Other classifiers were also tested during preliminary tests, but they were low in accuracy and thus not cited in our work. Table 1 presents the parameters for the top three classifiers.

As mentioned above, several combinations of data were used during the tests. Initially it was considered to train classifiers with a multiclass classification for the five stars available. However, the results in accuracy were around 57%, and it was then decided to reduce the number of classes. Next, 1 star was chosen for negative, 2, 3 and 4 stars for neutral and 5 stars for positive, leading to an accuracy of 73%. Finally, after considering previous research (Chang et al., 2017; Ghaddar & Naoum-Sawaya, 2018; Qiu, Liu, Li, & Lin, 2018) where the authors have used datasets for binary classification, it was decided to perform two tests using: *(i)* 1 and 2 stars for the negative class; 3, 4 and 5 for the positive, or *(ii)* 1, 2 and 3 stars for the negative class; 4 and 5 for the positive, with the results improving for the first combination of data. Table 2 shows the results for the classifiers in binary classification. Also, cross validation technique (v-fold cross validation with 10 folds) has been included in the experiments in order to prevent overfitting in the classifiers. The results obtained using or not cross validation are quite similar. Therefore, in order to reduce the computational complexity of the framework, the evaluation results detailed in Table 2 are for experiments without cross validation.

Table 1
Main parameters of machine learning algorithms used for the model.

System	Parameters
MLP	activation = logistic, alpha = 0.001, epsilon = 1e-08, hidden_layer_sizes=(100, 100), learning_rate = constant, learning_rate_init = 0.001, max_iter = 200, momentum = 0.9, power_t = 0.5, random_state = 1, shuffle = True, solver = adam
SVC	C = 1.0, cache_size = 200, coef0 = 0.0, decision_function_shape = ovr, degree = 3, gamma = auto, kernel = linear, max_iter=-1, probability = False, shrinking = True, tol = 0.001, verbose = False
LR	C = 1.0, dual = False, fit_intercept = True, intercept_scaling = 1, max_iter = 100, multi_class = ovr, n_jobs = 1, penalty = l2, solver = liblinear, tol = 0.0001, verbose = 0

Table 2
Evaluation results for binary classification (average for 10 runs).

System	Precision	Recall	F1-measure	Accuracy
MLP	0.88	0.88	0.88	0.882
SVC	0.88	0.88	0.88	0.878
LR	0.88	0.88	0.88	0.879
$LSVC_{l1}$	0.87	0.87	0.87	0.868
$LSVC_{l2}$	0.87	0.87	0.87	0.869
SGD_{l1}	0.86	0.86	0.86	0.862
SGD_{l2}	0.88	0.88	0.88	0.878
NB	0.86	0.86	0.86	0.860

5. Discussion

5.1. Machine learning model

The results show that the best classifiers for text categorization were MLP, SVC and LR. However, the algorithms were further tested eight times to compare the performance prediction accuracy. This is shown in Fig. 4, where the dots relate to the accuracy attained in each of the eight times the model was run.

Knowing which classifier best predicts positive and negative reviews

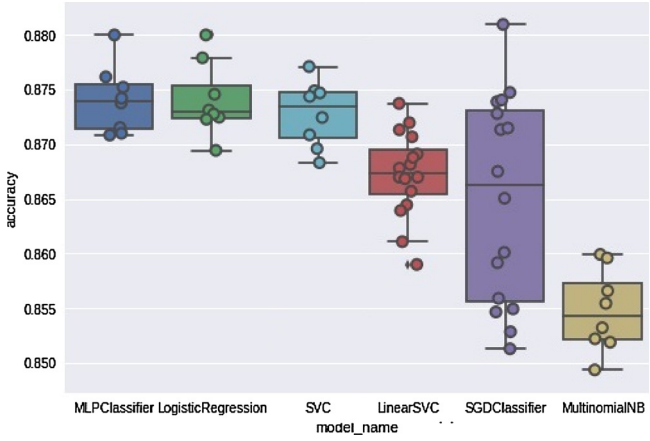

Fig. 4. Boxplot comparison of Machine Learning classifiers.

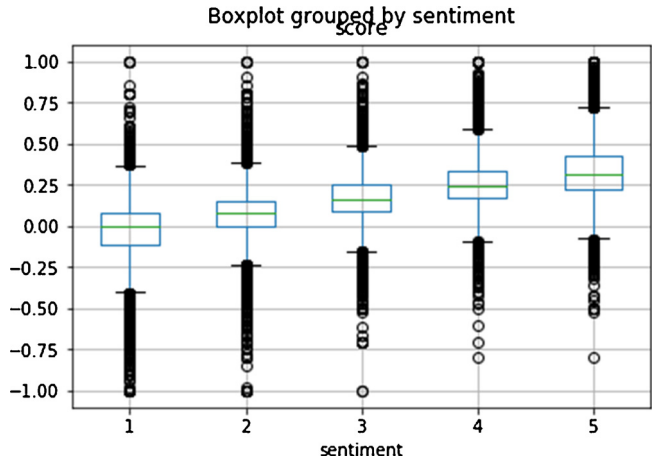

Fig. 5. Boxplot distribution of sentiment score for five classes of reviews.

allowed us to perform further data processing leading to insights that helped answer the questions posed earlier in this chapter. During preliminary tests with five-star classification, it was important to determine the sentiment score per category (stars), which is very helpful for processing the data in a quantitative manner. The boxplot in Fig. 5 shows the distribution of sentiment score for five classes, where 1 is absolute positive and -1 is absolute negative.

As can be seen in Fig. 5, reviews rated 1-star have a neutral sentiment score slightly skewed towards negative; as rating increases so does sentiment score. This figure also allows us to analyze the information for the qualitative aspects, since it could be interpreted as: users that rated 1 star have a rather neutral sentiment towards the business unit, but also that this sentiment is not that negative, indicating that these are areas of opportunity for improving customer service. On the other hand, reviews having a 5-star rating have a sentiment score in the positive area that barely reaches half of the absolute positive score, which also implies that many services could be improved in the area of customer care.

In regard to binary classification, Fig. 6 shows the distribution of sentiment score for one and five stars. This figure also allows us to process data for quantitative aspects: the representation of class 1 (one and two stars) is very similar to the representation of class 1 in Fig. 5, indicating, based on word weight, how similar the reviews for one and two classes can be. Then for class 5 (three, four and five stars), Fig. 6 shows that even though the grouping of these reviews has a positive sentiment score, the overall sentiment does not even reach half the positive score. Visualizing customers' mood this way will enable

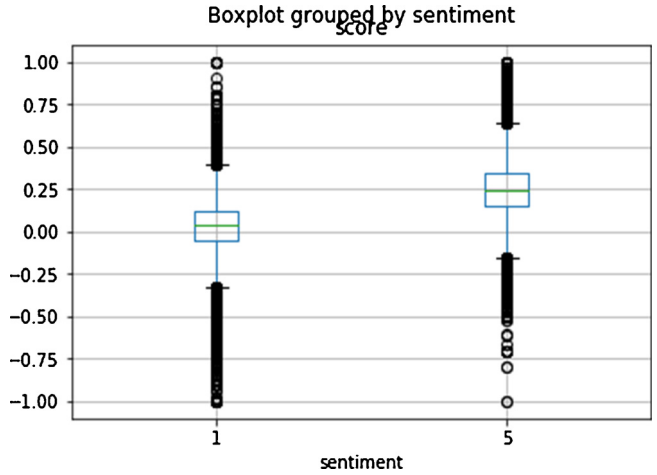

Fig. 6. Boxplot distribution of sentiment score for binary classification.

decision makers to take proper actions in a timely manner; this representation quantitatively shows how sentiment polarity is distributed if it is portrayed in a binary manner (absolute positive and negative).

To determine the main words that customers express in their reviews (the second question), the TFIDF coefficients of features were explored using the logistic regression classifier, which is presented in Fig. 7. These results show the main words that users express concerning services. In this figure, 0 represents neutrality, whereas negative values represent the most negative words mentioned across reviews; the right side of the figure shows the most positive words included in reviews. In other words, it could be said that Fig. 7 shows the analysis of the information both qualitatively and quantitatively. The words below the bars are the stemmed words in their root form.

Fig. 7 is considered by the authors as the one that shows more information regarding qualitative aspects of data. As seen on this figure, certain terms as "beauty", "great" or "love" reflect positive image of business as expressed by the users, whereas terms as "worst", "horribl" or "terribl" surely express disappointment when staying in a hotel or using certain services. The figure above shows the top of the best and the worst terms found in the corpus from a qualitative aspect, however during data experimentation other terms regarding specific circumstances as "air conditioned" or "parking lot" were found with lower TFIDF coefficients, these elements could be exploited in further research to discover more qualitative results.

These results show that by using the methodology proposed in the first stage of the framework, interesting insights can be obtained from unstructured heterogeneous data. Results obtained by the framework were able to process information both qualitative and quantitative. By integrating different elements, this paper has proven that companies and organizations can use ML to find hidden patterns in data that will allow them to gain competitive advantages to outperform competitors. Regarding the data used in this study, the results still apply to a variety of businesses in the tourism domain, which in turn belong to different cities; this issue could be considered somewhat misleading towards the outcome of the MLM. However, as mentioned above the shortage of public data (Jimenez-Marquez, Gonzalez-Carrasco, & Lopez-Cuadrado, 2018) to conduct these type of studies influences the sample size, this was a constraint for the train and test datasets used since a ML algorithm learns better when it has more data to learn from.

The authors consider that this does not limit the use of the framework for future research, the methodology proposed for the first stage can lead to similar or even better results for a more specific case of study when more data are available. Moreover, further analysis techniques could have been applied, and different results could be obtained from data, nevertheless this study is focused on presenting the framework and a case study based on its application, leaving for future research the integration of additional techniques to get results for different case studies. Ultimately, since there is not a total correspondence of the ML algorithms available for text analysis in BD and non-BD environments, the design of the ML algorithms selected to construct the model should be done with caution.

5.2. Framework discussion

The rapid time of development and the short time to market in today's world limit the time that data scientists have for data analysis in companies. However, every data analysis project should follow a methodology to ensure high quality results. The proposed framework follows a series of two stages divided into several phases for data analysis for BSD. The first stage can be done in a non-BD environment and the proposed framework highlights that a conceptual model must be created before analysis. The MLM constructed in the first stage is intended to serve as a base for the BD analysis expected to occur in stage two. This is because the two stages have the same purpose: to analyze a set of supervised data in different environments. The framework as a whole is not linked to a specific provider, methodology or ML

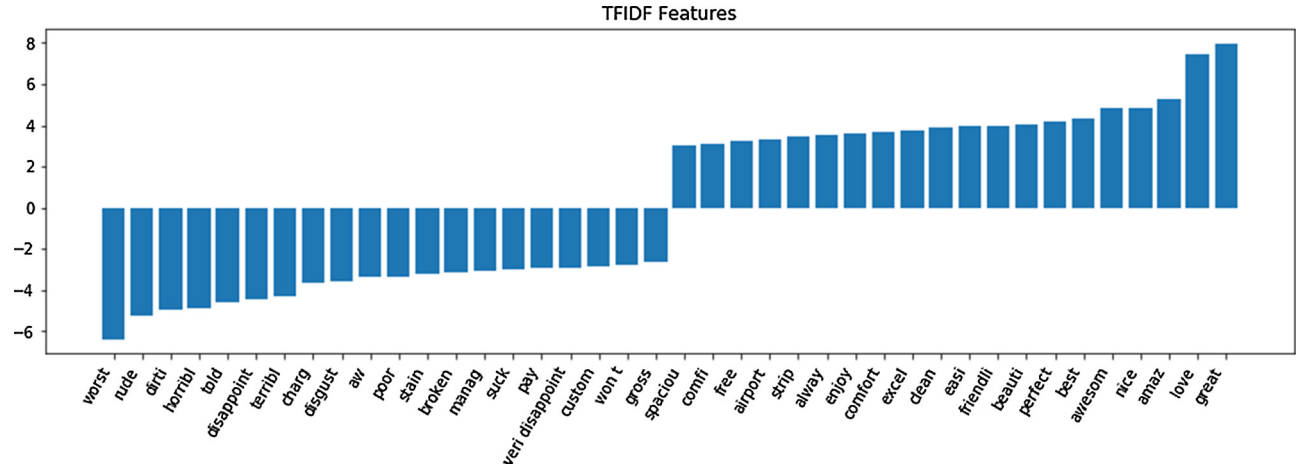

Fig. 7. Depiction of TFIDF features.

algorithm, which is what authors consider the proposal's strength regarding other BD frameworks in the existing literature, as mentioned in Section 2 and summarized next. With respect to the framework proposed by Vajirakachorn and Chongwatpol (2017), the steps these authors propose in their framework share many similarities with the phases proposed in this paper (set up objectives, collect data, analyze data by ML techniques, etc.), however, *(i)* their framework does not consider the analysis of textual data; (ii) the data management techniques proposed are related to business intelligence and database management tools, whereas the proposed in this paper does consider integrating BD tools and techniques; and (iii) their framework is intended to work with only structured data, while the proposed framework can analyze both structured and unstructured data. The framework proposed by Habib et al. (2016) has as its main purpose the 'big data reduction at the customer end'. Even though this framework considers more elements regarding the BD infrastructure and methodology with respect to the proposed framework, as previously mentioned, the aforementioned BD framework does not consider the integration of ML and NLP for data analysis to solve specific challenges when dealing with supervised data. The framework proposed by Yuan et al. (2016) considers tourist information crawled from travel blogs, and they "implement the frequent pattern mining method" to identify a city's popular locations by creating word vectors to construct a word network. Even though this framework works with tourist information extracted from reviews, it is limited to this domain, while the proposed framework is able to work with several domains.

Thus, the main theoretical implications of this research, regarding tourism domain, can be summarized in two main aspects. First of all, although other related frameworks could be found in the literature, they are usually linked to a unique technique to solve a particular problem or they are more complex (the internal structure of the framework is bigger). Therefore, in this case, the novelty resides in two fundamental aspects: (i) the processes to accelerate the delivery of results have been simplified, so there are fewer stages than in other related frameworks, and (ii) the techniques to solve analysis problems could be improved or even replaced in future research in order to improve the results, so the proposed framework is able to adjust to new conditions. Finally, considering the adaptability of the framework, it can be used a basis for future models for different problems and even domains. In this proposal, the capacity of the framework has been tested and explored with supervised data, but a future line of research could be to include data for different machine learning algorithms (unsupervised, reinforced, semi-supervised) since, in practice, the data obtained from Social networks is not tagged (Facebook, Twitter, etc.). Therefore, tourism companies could find value and extract knowledge from this social networks data if these solutions are explored.

In the practical side, companies in the tourism sector that have quantitative/qualitative data could take advantage of the data analysis features proposed in this research. Thanks to the adaptability of the framework, another type of tests could be performed by applying additional techniques in order to obtain new indicators for companies in the tourism sector. Finally, one of the most interesting practical implications lies on the fact that the proposed framework performs the data analysis in two stages, so that companies or organizations before deciding to mount or contract a Big Data architecture for the second stage, can carry it out a proof of concept in a smaller environment with the first stage. If favorable partial results are obtained, a big data architecture could be introduced in the company. Furthermore, this two-stages framework also allow to verify if the analysis of the data requires big data techniques depending on the volume of information to deal with and the hardware capabilities of the company.

6. Conclusions

The industry 4.0 is generating more data than ever before in the history of humanity. High-level techniques to store, manage, analyze and visualize data are being continuously created. However, this wide spectrum of possibilities to conduct a study or research using such a variety of elements makes it difficult to choose a certain development path to achieve the best results in the shortest time. Having a clear goal of what the expected results are, is the most important aspect to consider before conducting a study at this level. The framework is proposed to serve as a bridge between data analysis and technologies.

The most important aspects of the framework are: *(i)* it is a two-stage framework; the first is mainly focused on getting the data ready and finding the optimal MLM to analyze data; the second is concerned with setting up a BD infrastructure able to perform analysis and visualization tasks; *(ii)* with respect to other existing frameworks, the phases in the first stage have been reduced in order to get results in the shortest time by reducing complexity allowing the inclusion of new domains and algorithms; and *(iii)* the MLM is not related to specific ML algorithms, and therefore it increases the flexibility to use the framework since existing ML models can be incorporated to compare existing results. In this paper, results shown are limited to the first stage, and the data available is limited to the Yelp dataset. Not all the methods applied in the first stage can be scalable directly to big data analytics techniques. Since there is not a total correspondence of the ML algorithms available for text analysis in BD and non-BD environments, future research is centered on the application of the results of the first stage in the second one by means of BD specific ML algorithms and techniques. Future research should also be able to explore how to extend the framework through the integration of advanced machine learning

methods. Moreover, it is proposed to expand the research through a study that explores the differences between the two computation paradigms exposed (big data and not big data) to specify why these differences sometimes prevent the same methods from being used.

Current experiments have been limited to Yelp dataset. Future research can consider information from other sources (by agreements with social networks or by crawling public data). Also, future research will include variations in the algorithms applied in order to consider more sentiment categories as well as visual analysis of data. In the area of analysis, as mentioned in Section 4.5, the classification of reviews has been limited to two sentiment categories. Future research will expand the number of categories in order to consider five categories.

The analysis of how users rate services both quantitatively and qualitatively is an ongoing subject of study in computer sciences. Complex aspects in the subjectivity of text reviews and the many features existing in the datasets (price, location, opening hours, influencers, etc.) make data analysis in the BD era a permanent and evolving task. Only companies that integrate these methodologies into their business strategies can evolve and stay ahead of competitors in this data-driven age.

Declarations of interest

None.

References

Ahmad, S. N., & Laroche, M. (2017). Analyzing electronic word of mouth: A social commerce construct. *International Journal of Information Management, 37*(3), 202–213. https://doi.org/10.1016/j.ijinfomgt.2016.08.004.

Alahmadi, D. H., & Zeng, X.-J. (2015). ISTS: Implicit social trust and sentiment based approach to recommender systems. *Expert Systems With Applications, 42*(22), 8840–8849. https://doi.org/10.1016/j.eswa.2015.07.036.

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems With Applications, 77*, 236–246. https://doi.org/10.1016/j.eswa.2017.02.002.

Bayramusta, M., & Nasir, V. A. (2016). A fad or future of IT?: A comprehensive literature review on the cloud computing research. *International Journal of Information Management, 36*(4), 635–644. https://doi.org/10.1016/j.ijinfomgt.2016.04.006.

Carcillo, F., Dal Pozzolo, A., Le Borgne, Y. A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). SCARFF: A scalable framework for streaming credit card fraud detection with spark. *Information Fusion, 41*, 182–194. https://doi.org/10.1016/j.inffus.2017.09.005.

Chang, Y., Ku, C., & Chen, C. (2017). Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. *International Journal of Information Management,* (April), 1–17. https://doi.org/10.1016/j.ijinfomgt.2017.11.001.

Chen, L., Yan, D., & Wang, F. (2017). User perception of sentiment-integrated critiquing in recommender systems. *International Journal of Human Computer Studies,* 1–17. https://doi.org/10.1016/j.ijhcs.2017.09.005 000 (September).

Colace, F., De Santo, M., Greco, L., Moscato, V., & Picariello, A. (2015). A collaborative user-centered framework for recommending items in Online Social Networks. *Computers in Human Behavior, 51*, 694–704. https://doi.org/10.1016/j.chb.2014.12.011.

Costa, C., & Santos, M. Y. (2017). The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. *International Journal of Information Management, 37*(6), 726–734. https://doi.org/10.1016/j.ijinfomgt.2017.07.010.

Cybulski, J. L., Keller, S., Nguyen, L., & Saundage, D. (2015). Creative problem solving in digital space using visual analytics. *Computers in Human Behavior, 42*, 20–35. https://doi.org/10.1016/j.chb.2013.10.061.

Dar, S. (2016). A simulator for Spark scheduler. *Eindhoven.*.

Deng, N., & Robert, X. (2018). Feeling a destination through the "right" photos : A machine learning model for DMOs' photo selection. *Tourism Management, 65*, 267–278. https://doi.org/10.1016/j.tourman.2017.09.010.

Etaiwi, W., & Naymat, G. (2017). The impact of applying different preprocessing steps on review spam detection. *Procedia Computer Science, 113*, 273–279. https://doi.org/10.1016/j.procs.2017.08.368.

Etaiwi, W., Biltawi, M., & Naymat, G. (2017). Evaluation of classification algorithms for banking customer's behavior under Apache Spark Data Processing System. *Procedia Computer Science, 113*, 559–564. https://doi.org/10.1016/j.procs.2017.08.280.

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations – A case from Sweden. *Journal of Destination Marketing & Management, 3*(4), 198–209. https://doi.org/10.1016/j.jdmm.2014.08.002.

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

Ghaddar, B., & Naoum-Sawaya, J. (2018). High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research, 265*(3), 993–1004. https://doi.org/10.1016/j.ejor.2017.08.040.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management, 59*, 467–483. https://doi.org/10.1016/j.tourman.2016.09.009.

Habib, M., Chang, V., Batool, A., & Ying, T. (2016). Big data reduction framework for value creation in sustainable enterprises. *International Journal of Information Management, 36*(6), 917–928. https://doi.org/10.1016/j.ijinfomgt.2016.05.013.

Hu, Y. H., & Chen, K. (2016). Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. *International Journal of Information Management, 36*(6), 929–944. https://doi.org/10.1016/j.ijinfomgt.2016.06.003.

Iqbal, R., Doctor, F., More, B., Mahmud, S., & Yousuf, U. (2016). Big Data analytics: Computational intelligence techniques and application areas. *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2016.05.020.

Jimenez-Marquez, J. L., Gonzalez-Carrasco, I., & Lopez-Cuadrado, J. L. (2018). Challenges and opportunities in analytic-predictive environments of big data and natural language processing for social network rating systems. *IEEE Latin America Transactions, 16*(2), 592–597. https://doi.org/10.1109/TLA.2018.8327417.

Kubina, M., Varmus, M., & Kubinova, I. (2015). Use of big data for competitive advantage of company. *Procedia Economics and Finance, 26*(15), 561–565. https://doi.org/10.1016/S2212-5671(15)00955-7.

Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management, 36*(5), 700–710. https://doi.org/10.1016/j.ijinfomgt.2016.04.013.

Law, R., Rong, J., Quan, H., Li, G., & Andy, H. (2011). Identifying changes and trends in Hong Kong outbound tourism. *Tourism Management, 32*(5), 1106–1114. https://doi.org/10.1016/j.tourman.2010.09.011.

Li, J., Ott, M., & Varadarajan, B. (2013). *Identifying manipulated offerings on review portals. 2013 Conference on empirical methods in natural language processing (EMNLP)*.

Li, X., Pan, B., Law, R., & Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management, 59*, 57–66. https://doi.org/10.1016/j.tourman.2016.07.005.

Lin, A., & Chen, N. C. (2012). Cloud computing as an innovation: Percepetion, attitude, and adoption. *International Journal of Information Management, 32*(6), 533–540. https://doi.org/10.1016/j.ijinfomgt.2012.04.001.

Lin, X., Li, Y., & Wang, X. (2017). Social commerce research: Definition, research themes and the trends. *International Journal of Information Management, 37*(3), 190–201. https://doi.org/10.1016/j.ijinfomgt.2016.06.006.

Lismont, J., Vanthienen, J., Baesens, B., & Lemahieu, W. (2017). Defining analytics maturity indicators: A survey approach. *International Journal of Information Management, 37*(3), 114–124. https://doi.org/10.1016/j.ijinfomgt.2016.12.003.

Marrese-Taylor, E., Velásquez, J. D., Bravo-Marquez, F., & Matsuo, Y. (2013). Identifying customer preferences about tourism products using an aspect-based opinion mining approach. *Procedia Computer Science, 22*, 182–191. https://doi.org/10.1016/j.procs.2013.09.094.

Mueller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python* (First Edit). O'Reilly Media, Inc.

Ngo-Ye, T. L., & Sinha, A. P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems, 61*(1), 47–58. https://doi.org/10.1016/j.dss.2014.01.011.

Olmedilla, M., Martínez-Torres, M. R., & Toral, S. L. (2016). Harvesting Big Data in social science: A methodological approach for collecting online user-generated content. *Computer Standards & Interfaces, 46*, 79–87. https://doi.org/10.1016/j.csi.2016.02.003.

Oussous, A., Benjelloun, F., Ait, A., & Belfkih, S. (2017). Big data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences,* 1–18. https://doi.org/10.1016/j.jksuci.2017.06.001.

Özköse, H., Sertac, E., & Gencer, C. (2015). Yesterday, today and tomorrow of big data. *Procedia - Social and Behavioral Sciences, 195*, 1042–1050.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2.

Pranata, I., & Susilo, W. (2016). Are the most popular users always trustworthy? The case of Yelp. *Electronic Commerce Research and Applications, 20*, 30–41. https://doi.org/10.1016/j.elerap.2016.09.005.

Qiu, J., Liu, C., Li, Y., & Lin, Z. (2018). Leveraging sentiment analysis at the aspects level to predict ratings of reviews. *Information Sciences, 451–452*, 295–309. https://doi.org/10.1016/j.ins.2018.04.009.

Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management, 38*(1), 187–195. https://doi.org/10.1016/j.ijinfomgt.2017.07.008.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis : Tasks, approaches and applications. *Knowledge-based Systems, 89*, 14–46. https://doi.org/10.1016/j.knosys.2015.06.015.

Reuther, A., Byun, C., Arcand, W., Bestor, D., Bergeron, B., Hubbell, M., ... Kepner, J. (2018). Scalable system scheduling for HPC and big data. *Journal of Parallel and Distributed Computing, 111*, 76–92. https://doi.org/10.1016/j.jpdc.2017.06.009.

Rong, J., Quan, H., Law, R., & Li, G. (2012). A behavioral analysis of web sharers and browsers in Hong Kong using targeted association rule mining. *Tourism Management, 33*(4), 731–740. https://doi.org/10.1016/j.tourman.2011.08.006.

Scikit-learn CountVectorizer (2018). *CountVectorizer*. Retrieved fromhttp://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

Scikit-Learn TFIDFVectorizer (2018). *TfidfVectorizer*. Retrieved fromhttp://scikit-learn.

org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.
Senyo, P. K., Addae, E., & Boateng, R. (2018). Cloud computing research: A review of research themes, frameworks, methods and future research directions. *International Journal of Information Management, 38*(1), 128–139. https://doi.org/10.1016/j.ijinfomgt.2017.07.007.
Shirdastian, H., Laroche, M., & Richard, M. O. (2017). Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. *International Journal of Information Management,* (August), 0–1. https://doi.org/10.1016/j.ijinfomgt.2017.09.007.
Silva, T. C., & Zhao, L. (2015). High-level pattern-based classification via tourist walks in networks. *Information Sciences, 294*, 109–126. https://doi.org/10.1016/j.ins.2014.09.048.
Sultan, N. (2013). Cloud computing: A democratizing force? *International Journal of Information Management, 33*(5), 810–815. https://doi.org/10.1016/j.ijinfomgt.2013.05.010.
Tan, W., Blake, M. B., Saleh, I., & Dustdar, S. (2013). Social-network-sourced big data analytics. *IEEE Internet Computing, 17*(5), 62–69. https://doi.org/10.1109/MIC.2013.100.
Vajirakachorn, T., & Chongwatpol, J. (2017). Application of business intelligence in the tourism industry: A case study of a local food festival in Thailand. *Tourism Management Perspectives, 23*, 75–86. https://doi.org/10.1016/j.tmp.2017.05.003.
Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management, 58*, 51–65. https://doi.org/10.1016/j.tourman.2016.10.001.
Xu, Y., & Yin, J. (2015). Collaborative recommendation with user generated content. *Engineering Applications of Artificial Intelligence, 45*, 281–294. https://doi.org/10.1016/j.engappai.2015.07.012.
Xu, X., Wang, X., Li, Y., & Haghighi, M. (2017). Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. *International Journal of Information Management, 37*(6), 673–683. https://doi.org/10.1016/j.ijinfomgt.2017.06.004.
Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management, 36*(6), 1231–1247. https://doi.org/10.1016/j.ijinfomgt.2016.07.009.
Yelp (2016). *Yelp dataset challenge.* Retrieved fromhttps://www.yelp.com/dataset/challenge.
Yuan, H., Xu, H., Qian, Y., & Li, Y. (2016). Make your travel smarter: Summarizing urban tourism information from massive blog data. *International Journal of Information Management, 36*(6), 1306–1319. https://doi.org/10.1016/j.ijinfomgt.2016.02.009.