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

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