CONTENT MINING FRAMEWORK IN SOCIAL MEDIA: A FIFA WORLD CUP 2014 CASE ANALYSIS

ABSTRACT: This paper proposes a social media content mining framework that consists of seven phases, namely: Objectives, domains, and other definitions; Technology tools and support definition; Data Collection Strategies and Execution; Data Cleaning and Treatment; Data Mining; Results Interpretation and Evaluation; Results Visualization.. The framework was tested empirically during the FIFA World Cup 2014 at Curitiba (Brazil) as one of the main host city destinations. The research focused on mining of Twitter content with tourist services ontology (hospitality, food & beverages and transportation). In total 58,686 valid messages were collected, analyzed and associated with an application ontology. Content analysis demonstrated an accurate, real time reflection of tourism services. The framework is effective to collect relevant content and identify popular topics in social media towards strategic and operational tourism management. A better understanding of dynamic conversations facilitates better and faster decision making and a rapid response in real time that can support tourism organization and destination competitiveness.

Keywords: Social Media; Content Mining; Twitter; Tourist Services; Brazil; FIFA World Cup 2014.

1. Introduction

Social media is having a great impact on travel (Xiang; Magnini; Fesenmaier, 2015, Buhalis and Law, 2008). These forms of online communication are empowering travelers to not only interact with business of the tourism industry, but also to exchange opinions with others online (Xiang; Wang; O'Leary; Fesenmaier, 2015). Social media can generate electronic versions of traditional word-of-mouth, and often consist of comments published by travelers on the tourism products/services they experience (Filieri; Mcleay, 2014, Williams, Inversini, Buhalis, Ferdinand, 2015). Also, there is an awareness among tourism managers that social media content and reviews form a rich source of data (Phillips; Zigan; Silva; Schegg, 2015).

The popularity of social media has been widely recognized by studies that highlight the preference and search for information on reviews posted online by other tourists (Gretzel; Yoo, 2008; Liu; Park, 2015; Xiang; Gretzel, 2010). In a study conducted by Expedia Media Solutions in partnership with comScore (Expedia Media Solutions; ComScore, 2013) it was revealed that 43% of respondents published reviews and travel-related content on social media. A recent study on how travelers used review sites while planning their holiday in the United Kingdom (Statista, 2016) showed that 38% of the respondents use online review sites always or for most of the holidays taken, and 30% are occasional users. These figures indicate that reviews can be a valuable source of useful data as a key component of smart tourism destinations management (Marine-Roig; Clavé, 2015). The potential of social media to affect markets by driving consumer choice can have a discriminating effect on the tourism industry (Phillips; Zigan; Silva; Schegg, 2015, Leung, Law, van Hoof, Buhalis, 2015).

Monitoring what is being said online about the destination, product, service or tourism organization for consumers and tourists in social media offers enormous opportunities and benefits. For example: (i) Relationship and user engagement (Hea; Zha; Li, 2013; Paine, 2011); (ii) Competitive analysis (Hea; Zha; Li,

2013 ; (iii) Sentiment and opinion analysis (Pol; Patil; Patankar; Das, 2008; Han; Kamber; Pei, 2012); (iv) Knowledge discovery (Hea; Zha; Li, 2013; Han; Kamber; Pei, 2012; (v) Consumer knowledge management (Hea; Zha; Li, 2013; Chua; Bannerjee, 2013; Chua, 2011; Magnier-Watanabe; Yoshida; Watanabe, 2010) ; (vi) Management and decision making process (Han; Kamber; Pei, 2012; Buhalis and Foerste, 2015, Lau, Lee, Ho, 2005;); (vii) Social Media Strategy (Hea; Zha; Li, 2013); (viii) Prediction of scenarios, trends and events (Hea; Zha; Li, 2013; Kalampokis; Tambouris; Tarabanis, 2013; Yu; Kak, 2012); (ix) Creation and innovation of products and services (Chua; Bannerjee, 2013; Chua, 2011). Such activities are interconnected and support each other to get results and achieve the objectives of social media mining activities. Figure 1 illustrates these activities, the benefits and opportunities.

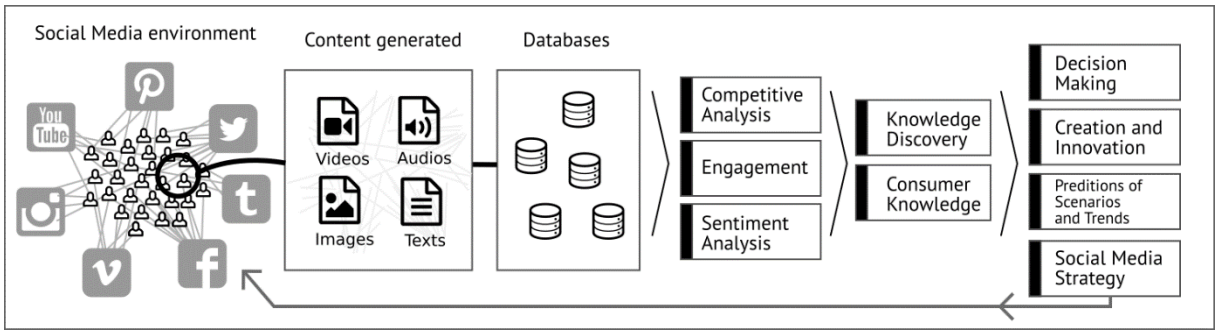

FIGURE 1: OPPORTUNITIES AND BENEFITS IN SOCIAL MEDIA CONTENT MINING

Research addressing the content mining theme in social media is gaining more popularity. Kalampokis, Tambouris and Tarabanis (2013) investigated the use of data from social media for predictions in areas such as disease outbreaks, volatility in the stock market and election results. While Abrahams, Jiao, Fan, Wang and Zhang (2013) used text mining techniques to detect consumer complaints about automobiles. Although the importance of social media in the travel industry has been an emerging research topic (Leung et al, 2013; Amaro, Duarte, Henriques, 2016), the content mining approach is at an early stage in tourism. For instance, He, Zha, and Li (2013) used the text mining to analyze content in profiles of Facebook and Twitter of the three largest US pizza chains. Neves and Marchiori (2014) monitored Twitter content to evaluate the perceived quality of tourism products and services during the London 2012 Olympic Games. Also, Williams, Inversini, Buhalis and Ferdinand (2015) analyzed the interactions inside communities of interest on Twitter during Bournemouth AirShow2013 for a better understand of online word-of-mouth (eWOM) statements.

The goal of this study is to propose and test a framework for content mining in social media. The framework consists of seven phases and was tested during the FIFA World Cup 2014, using Curitiba (Brazil). Curitiba was selected for being one of the main tourist destinations in Brazil, internationally recognized and awarded for its urban planning and management, environment, public transport initiatives. It is also famous for its rich ethnic and cultural diversity a result of the large number of European immigrants throughout the twentieth century, represented in various memorials, parks, squares and other tourist attractions of the largest city in southern Brazil. Considering the potential of big events to attract people's attention and the fact that Curitiba was one of the host cities during the FIFA World Cup held in Brazil, it was an opportunity to monitor the content published on social media, about the city, by users from other states and countries with different cultures and experiences.

Content mining was held on Twitter data, followed by testing empirically the stages established in the framework. In total, 58,686 valid messages were collected, analyzed and associated with an application ontology. This study amalgamates concepts, discussions, models, techniques and tools from other fields, such as computer science, data mining, big data, machine learning and social network analysis and makes a methodological contribution to the tourism literature. The framework and its application can be used as a decision support system tool for management, promotion, development and planning of tourism by Destination Management Organizations (DMO). In addition, Smart tourism destination programs can incorporate this framework as part of their ICT component support (Boes, Buhalis, and Inversini, 2015) that promotes dynamic and real time decision support and agile management.

2. Social Media and Tourism

Mistilis, Buhalis and Gretzel (2014) state that the management of a tourist destination is a challenge due to the complexity of a required relationship between numerous public and private actors involved in tourism. Lack of interaction among key actors can result in a competitive disadvantage (Blumrodt; Palmer, 2013). Therefore, tourist organizations must be present and connect with consumers and all stakeholders regularly, using a range of social media resources and platforms (So, King, Sparks, Wang, 2016). Neuhofer, Buhalis and Ladkin (2014) expands this discussion, remarking the tourist role in a co-create online experience through ICTs, and revealing a six dimensions' model of social intensity continuum, from disconnection to social co-living of the experience.

Social media are gaining prominence as a critical element of DMO strategy. As public sector budgets are reduced, managers are often forced to seek alternatives in the way marketing budgets are spent more effictively. Social media empower DMOs to reach and interact with their global audience with limited resources (Hays, Page, Buhalis; 2013). These tools have also become a new method for interacting with development agent and tourist information suppliers, concurring with DMOs and their role as provider of knowledge about destinations (Xiang; Gretzel, 2010).

Social media offer the opportunity to achieve a long-term strategic understanding whilst also offering a short-term tactical level of action, focused on real time and instance response to events. That is one of major differences of social media relative to previous sources of customer feedback / opinions / feelings. Contextualising and personalizing the tourism product adds considerable value to customer interaction with their environment and allows them to cocreate their experience. Buhalis and Foerste (2015) use the term SoCoMo to define a combination between social media, context-aware marketing and mobile devices, related to an advanced form of contextual (and custom) marketing totally based on a real time information/response to customer potential interests. Therefore, social media mining can provide opportunities to identify the views and feelings of tourists on destinations, products, tourist services, to monitor and act based on events and everyday situations and find weaknesses, strengths and opportunities in real time. Information shared on social media is by definition public and therefore access and analysis should available regardless of privacy legislation in different countries.

Social Media Mining (SMM) is a subject derived from Web Content Mining (extraction of knowledge from multimedia data on the Web such as images, videos and audio) using associated textual data (Han; Kamber; Pei, 2012) and Text Mining (models, trends, patterns, interesting relationships or useful and relevant rules from textual unstructured data extraction) (Hea; Zha; Li, 2013). Social media comprises of interactive platforms that facilitate communication, creation and sharing of content generated by users on internet such Twitter and Facebook (Kaplan; Haenlein, 2010; Chua; Bannerjee, 2013; Zeng; Gerritsen, 2014). The development of tools and application techniques to collect, monitor, summarize and analyze social media content facilitates Social Media Mining (SMM) (Zafarani; Abbasi; Liu, 2014). SMM consists of three general steps, namely: (i) data extraction providers and social media servers through application programming interfaces (API); (ii) analysis, data integration and storage; and iii) analysis of data to extract information of interest (Crooks; Croitoru; Stefanidis; Radzikowski, 2013). SMM application offers opportunities to discover and identify patterns, features, information and relevant topics; give interesting perspectives for the understanding of human behavior; perform qualitative and quantitative analysis and even predict future events from unstructured content requiring. The key advantage of social media is that data can be collected in real time facilitating a rapid response and enabling organisations and destinations to cocreate tourism products dynamically. Permission marketing will effectively be the response to different legislations which may protect consumers from retargeting.

Collecting data from social media is the major challenge in SMM. The largest portion of the content is written ignoring spelling and grammar rules; featuring lexical and syntactic problems such as slang, abbreviations, words settings, use of *emoticons*, creating new words, multiple meanings, among others (Paine, 2011; Abrahams et al, 2013). In addition, cultural differences can influence the different ways of expressing sentiments such as metaphors and other linguistic patterns, increasing the challenge for SMM. When emerging characteristics are modeled, SMMs also can help on detecting topics and identifying them in real time as pointed by Chen, Amiri, Li and Chua (2013). Paine (2011) suggests to use hashtags as a reference to be monitored as a representation of an event or subject. High incidence of spam messages on recovery is also a known challenge, as showed by Paine (2011): on average from 85% to 95% of all content collected and approximately 70% irrelevant.

Although various tools and websites permit the execution of SMM phases, in most cases, the demand for strategic information has such a degree of specificity that generalization of available patterns is not sufficient, requiring the development of systems (Crooks et al., 2013). Therefore, SMM techniques and tools to collect, share, explore and visualize social media data have been extensively explored and developed (Tang; Yang, 2012). Human efforts are also considered, on the creation of ontologies to support retrieval, fill preferences and marking documents (manual modeling) or assisting computational techniques (automatic modeling, semi-automatic modeling) (Chanana, Ginige, Murugesan, 2004). Ontologies are used as content theories about the types of objects, properties of objects and relationships between objects that are possible in a particular knowledge domain (Chandrasekaran, Josephson, 1999).

Social Media Mining (SMM) Methods

A variety of methods have been developed to investigate, collect, analyze content, feelings and topics in social media. These methods require interdisciplinary skills as they involve raw data transformation into

meaningful information and knowledge, applying the most appropriate techniques and methods to each situation (Abrahams et al, 2013; Zafarani, Abbasi; Liu, 2014). Technical categorization of unstructured data is adapted and applied in social media studies. For example, Chen et al. (2013) showed that keyword-based approaches and high frequency terms can be a good indicator for topics in social media. Clustering or grouping techniques offer the advantage of revealing unforeseen patterns (Chen; Liu, 2004). The problem itself will define the best approach to the researchers and analysts (Barbier; Liu, 2011). The following methods were selected for further analysis in order to support the development of the framework.

Method proposed by Hea, Zha, and Li (2013)

Hea, Zha and Li (2013) conducted an in depth case study and developed a text mining method to analyze content and unstructured information from Facebook and Twitter profiles for the top three US pizzerias. The study sought to answer patterns found from profiles, and what are the main differences between the two social media. Figure 2 shows the main steps and activities.

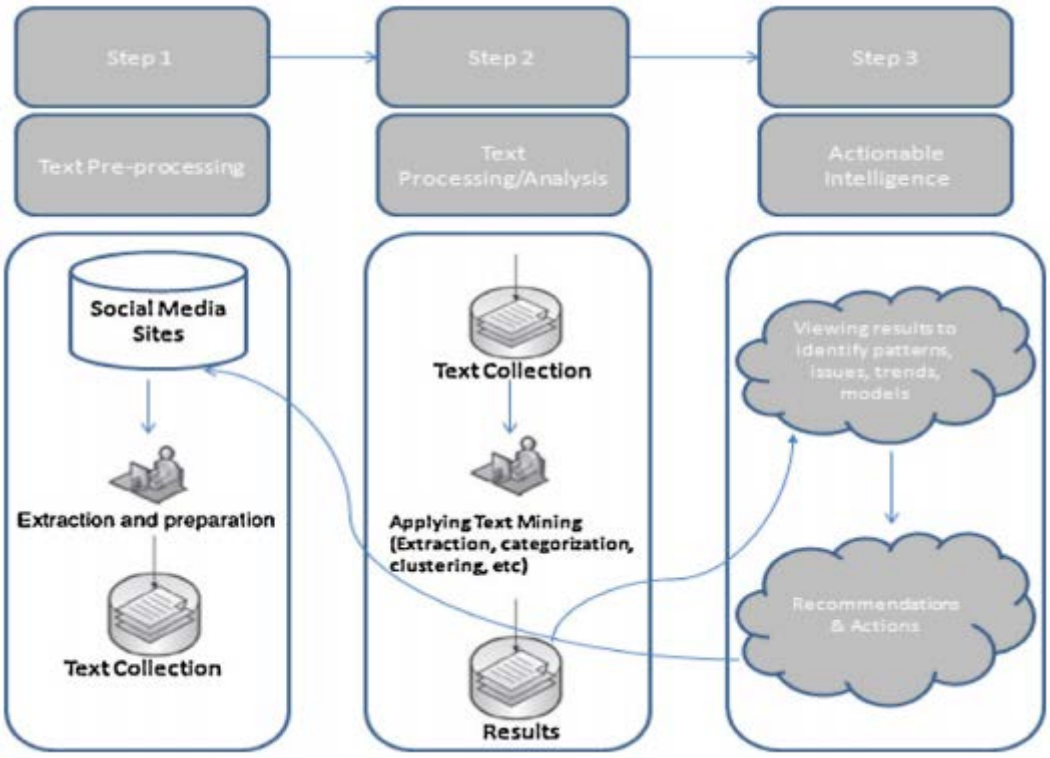

FIGURE 2: TEXT MINING PROCESS FOR SOCIAL MEDIA CONTENT (HEA; ZHA; LI, 2013)

The study consisted of two phases: (i) quantitative data were collected manually and individually through social media profiles, followed by pre-processing step of cleaning data; then (ii) text mining technique were applied to analyze content published. Three tools were used: Spreadsheets (initial structuring), Clementine (to explore and extract key concepts, create categories) and, NVivo (to perform an exploratory approach, testing ideas, finding interesting paths, etc.). The code process takes a lot of time and effort, for categories / dimensions creation and combination and validation procedures. Hea, Zha and Li (2013) identified and grouped Twitter posts in five different categories according to the subject of the tweets (Figure 3).

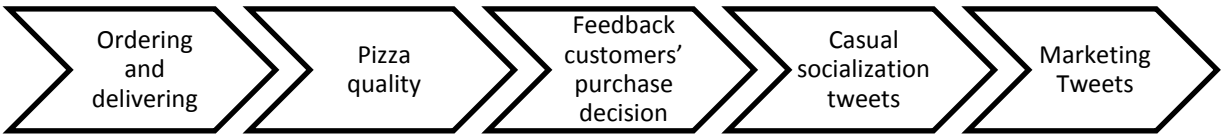

FIGURE 3: MAJOR CATEGORIES IDENTIFIED ON TWITTER CONTEXT OF US PIZZA CHAINS, ORDERED BY THE CUSTOMER EXPERIENCE (HEA; ZHA; LI, 2013)

The highlight of the methodology developed by Hea, Zha, and Li (2013) was the practical application in tourism to the area of food & beverage and two social media. It also showed the need of combining techniques and tools to respond a problem and have a comprehensive understanding of the issues discussed on social media.

Method proposed by Neves and Marchiori (2014)

The Neves and Marchiori (2014) method focused on the use, collection and interpretation of data and content published on Twitter. The authors applied their method on content analysis of perceived quality of tourism during the Olympic Games in London 2012, based on user posts. The data were collected and later analyzed according to a domain ontology with the following categories: food, lodging, transportation and security. The Microsoft ™ Analytics application was used to get data (n = 9,710) from Twitter based on the following criteria: "# (event name)" (ex.: "#london2012").

A sample of 4,440 Twitter messages were selected according to: use of English; use of Boolean *"and"* followed by specific research item; obligatory presence of the topic "# (event name)"; publication held during the official period of the event; excluding re-tweets. Of this total, a stratified sample of 878 tweets was clipped to undergo content analysis using the Bardin Assessment Analysis (2011). The benefit of the methodology developed by Neves and Marchiori (2014) was the practical application in tourism with a domain ontology and the use of hashtag strategy (#) and Boolean search of the items defined by the author to assist in targeting relevant and appropriate messages to the research objectives.

Method proposed by Kalampokis, Tambouris and Tarabanis (2013)

Kalampokis, Tambouris and Tarabanis (2013) aimed to raise the existing knowledge and provide a deeper understanding of the use and predictive power of social media applied in empirical studies. They looked for forecasts in diverse areas such as outbreaks diseases, product sales, results of elections and financial market volatility and stock on the stock exchange. The authors synthesized 52 articles in a conceptual framework of social media data analysis for forecasting that allowed to classify and evaluate existing knowledge (Figure 4).

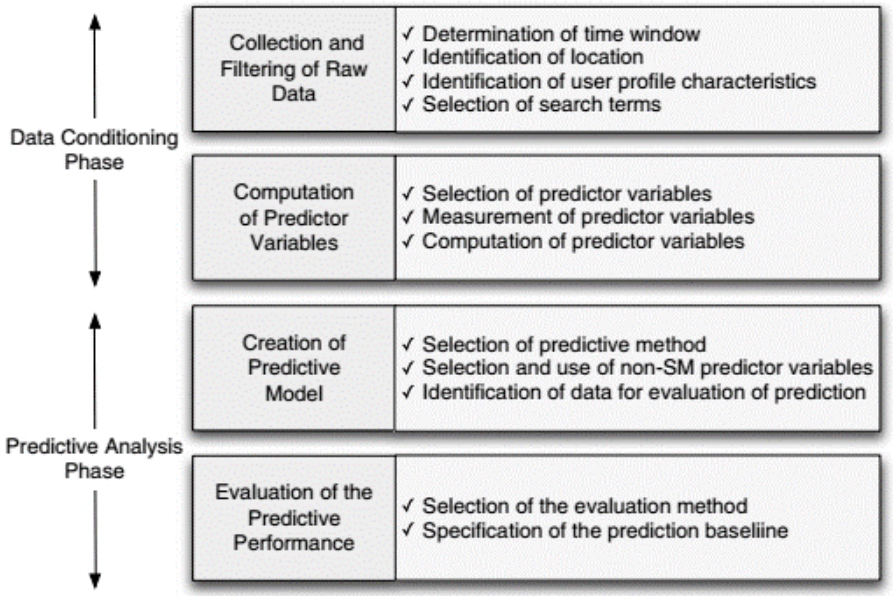

FIGURE 4: DATA MINING METHOD IN SOCIAL MEDIA (KALAMPOKIS, TAMBOURIS, TARABANIS, 2013)

The highlight of this method was the analysis of the existing knowledge, providing a deeper understanding of the use of social media data for predictions, as the two main phases required: (i) data conditioning and (ii) predictive analysis. This approach reinforce the need of a dynamic base that could support, a long-term analysis based on tests in data conditioning. As a consequence, the role of a self-response system can help with real time reactions, once a predictive model can show strategic findings faster than an exploratory model.

Method proposed by Abrahams, Jiao, Fan, Wang and Zhang (2013)

Abrahams et al. (2013) investigated, developed and implemented a social media text-mining model with content applied to the automotive industry. The model (Figure 5) can automatically identify parts that are subjects of publication from costumers in discussion forums. This enables the identification of the most significant terms of each component category, as well as the most popular topics.

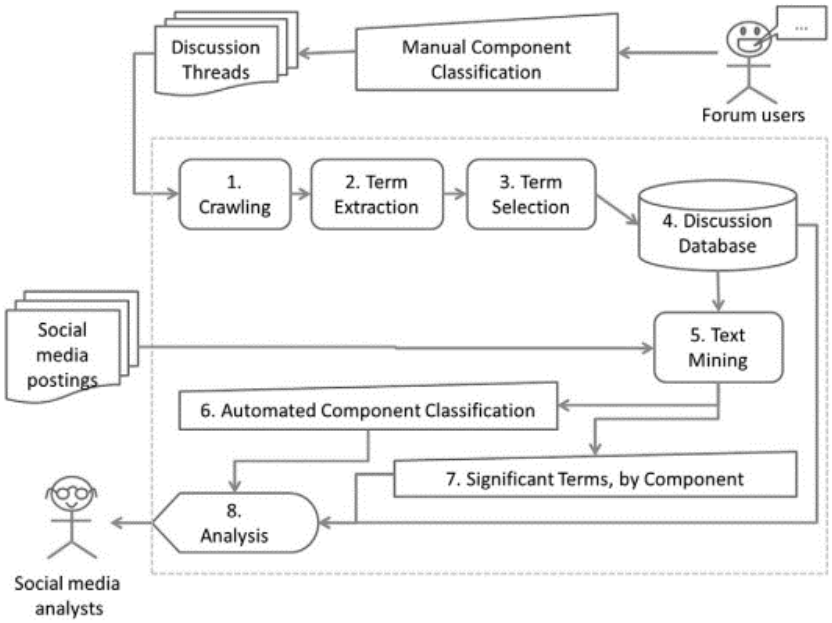

FIGURE 5: DATA MINING METHOD IN SOCIAL MEDIA PROPOSED BY ABRAHAMS et al. (2013)

The model starts with data gathering from discussion forums, followed by a relevant term extraction. After that, a term selection related to a category of analysis proceeds, before the database records the results. This is followed by a text mining application based on an automated component classification, when significant terms of each component /category are organized and classified. Analysts can visualize, for example, threads on issues and safety equipment related problems, regarding the performance of the vehicle, mapping and detecting the main subjects in each category.

Methods Overview

Each of the above methods provide a range of unique features and characteristics that allow SMM in depth. Combining the different methods and their strength enable us to synthesize and amalgamates all useful features. Hea, Zha and Li (2013) opted for manual categorization of messages, focusing the accuracy in the classification rather than the volume. Instead of using a set of categories, aimed to find them among the data. Neves and Marchiori (2014) in the investigation of the London Olympic Games also did manual categorization of messages using sampling to reduce the data size analysis. Unlike the previous model, the categories were defined before the collection, based on the structure of tourist services and the relevant messages were recorded accordingly. The collection was responding to a structure already formed on the data, unlike Hea, Zha and Li (2013) that sought to identify the structure itself. Hashtags were used to sample and although they revoke more relevant messages, they could also exclude opinions that may have been issued in other sentences.

Abrahams et al. (2013) also used an ontology to identify the relevant terms in online forums. Unlike Hea, Zha and Li's (2013) and Neves and Marchiori's (2014) models, the system was put into operation in real time, with automatic classification component treating the data as they were published, with higher gain scale and less use of time. Kalampokis, Tambouris and Tarabanis (2013) did not do an empirical study, but put in focus predictive models, i.e., a step beyond previous models to strengthen the possibility of predicting future events with historical data and not just summarize them.

Understanding these four models illustrates the processes and technologies required to support the SMM objectives. The data collection using hashtags can reduce the analysis work, but there is still a great volume of data to be analyzed and there may be content missing. Manuals categorization strategies as part of content analysis, can take a long time and possibly jeopardise reaction to short time decisions. The combination of categories of analysis (such as ontologies), automatic categorization and identifying trends (such as text mining) means an increase in scale and possibility of faster response. The analysis of previous methods allows a better understanding of the steps, processes and stages as well as problems to be faced and possible solutions, as well as the activities required to achieve research objectives.

Text mining techniques support the analysis of any message and, therefore, it is oriented towards classification of the message to data types. In turn, SMM explores where data is located in the social media sphere and elaborates the extraction of information. This can involve not only messages but also other types of data such as video, photo, and relationships between actors. Added to these areas, there is still content analysis cited in the study of Neves and Marchiori (2013), whose orientation is more qualitative and often not automated. Understanding the challenge posed in this study, the combination of these four sources of methods/techniques improves our understanding of content mining in social media. The study is not exhaustive in those fields, but combining them allows to develop a comprehensive approach as well as to propose and empirically test the proposed SMM model. The processes, steps and phases were extracted, combined and provide the basis for the construction of the proposed SMM model to assist DMOs in tourism and destination management.

3. Content Mining Framework in Social Media

Since the growth of social media, marketers around the world explore methods of accessing information, thoughts, media, sentiments, in real time. SMM approaches are still in early stages and although great progress has been made, there are still many open questions to be solved. This research area has been the focus of many researchers and therefore the methods proposed are increasing rapidly. However, they are still far from established concepts that can be applied globally, because social media data are also emerging rapidly and they are usually large, noisy, unstructured, and dynamic (Hea; Zha; Li, 2013; Barbier; Liu, 2011). The four methods in SMM detailed in the literature review have brought interesting points to contribute to the creation of content mining framework in social media, including:

- Importance of defining the scope, the concepts domain, and the problem to be solved from that mining;
- Definition of a technology support required (e.g., tools, techniques) to handle the data size, format, origin, structure efficiently, since data collection until presentation (between every step);
- Definition, testing, implementation and monitoring of different data collection strategies;
- Attention to data cleaning and processing in the format of a database with relevant content;
- Choosing one or more appropriate methods and performance of data mining;
- Summarizing results analyzing the adherence and the existence of patterns into categories;
- Defining a more suitable format for results visualization, meeting the demand of the user of such knowledge in the solution of the proposed problem.

Therefore, a content mining framework in social media was developed based on the methods studied. Key processes are listed below (Figure 6) and include: Objectives, domains, and other definitions; Technology tools and support definition; Data Collection Strategies and Execution; Data Cleaning and Treatment; Data Mining; Results Interpretation and Evaluation; Results Visualization. All the seven phases of the framework are explained and discussed in the following sections. The framework was developed and tested empirically during FIFA World Cup 2014, using Curitiba (Brazil) as a destination. Twitter was chosen as the social media source and given ethical and user privacy issues, the study only monitored public messages, that users shared for publications to be public in their privacy settings. The empirical results obtained from 58,686 tweets about tourist services in Curitiba are presented and discussed in section 4.

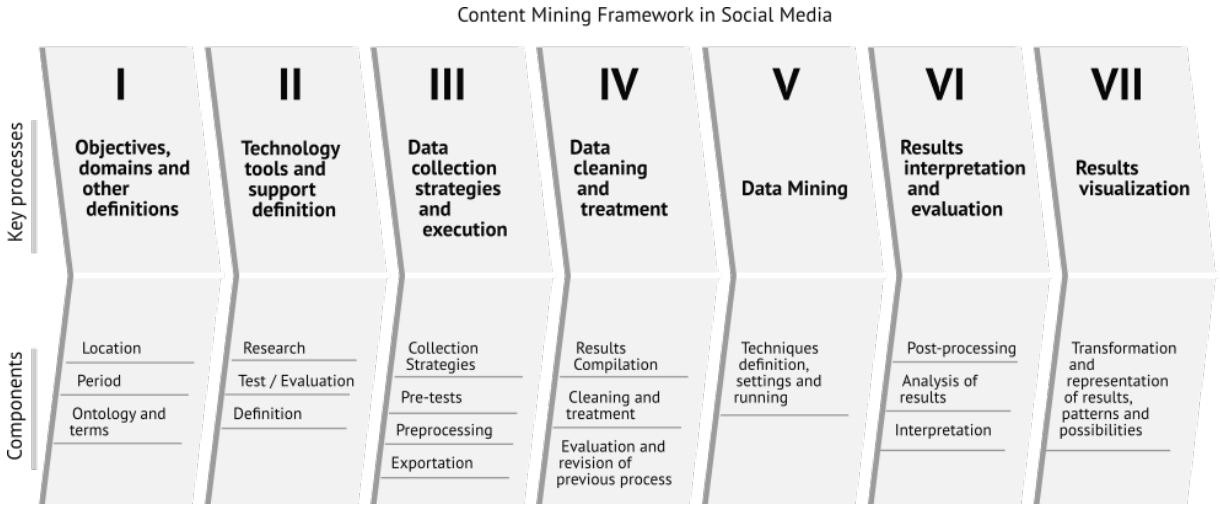

FIGURE 6: CONTENT MINING FRAMEWORK IN SOCIAL MEDIA

3.1. Objectives, domains, and other definitions

The first phase in SMM is to define the goals of monitoring and content mining, as well as the preparation of the data collection. Based on the Kalampokis, Tambouris and Tarabanis (2013) model, the process consists of three steps where the relevance of the collected content depends on four questions: Where, When, Who and What. The steps of identifying the user profile features concerning the Who question were not included in this study because its object is the textual content of social media posts and not users and authors characteristics.

The location step corresponds to defining Where content is collected. This can be translated as the social media content resources, selected in this research as Facebook groups and public posts, Twitter twits (messages) and YouTube videos. Sometimes geotagging is already included in the content available on social media whilst on other occasions location is evident in the content. For example some one may have done a check in Paris or the location is obvious because there is a recognizable landmark in the frame, such as the Aiffel Tower. Location inferences are of great importance as they drive contextual services (Buhalis and Foerste , 2015). The period was related to When and time tagging monitoring the timing of the content. Therefore, following Paine (2011), monitoring is a comparative strategy that requires result comparisons between different events and periods to extract value, meaningful information and knowledge. The present study was designed using five data collection milestones (from September 2013 to August 2014, in total).

Next step is a definition of research terms and languages. Selected terms are based on keywords, categories and characteristics defined according to the matters, topics of interest and specific objectives of the monitoring. At this stage, an application ontology should be designed to enable and identify rules of associative relationships that facilitate not only the ontology development itself but also serve as a basis for performing logical, organized and efficient searches (Amorim; Cheriaf, 2007).

Neves and Marchiori (2014) proposed four categories (Food and Beverage; Accommodation; Transport; Security) and forty terms (in English) on their domain ontology. Instead of the related goals, some adjustments were needed to meet the aims of this study. The category "Security" was excluded because this study focused on tourist services and attractions, so the category “Tourist Attractions” was created. The scope of terms has been expanded to three languages: Portuguese, English and Spanish and the number of terms was revised to expand some issues not addressed by Neves and Marchiori (2014).

The application ontology is used from a custom domain approach, not the OWL (Ontology Web Language); a decision based on a lower need of complexity and a specific application, rather than the purpose of OWL. Furthermore, from the list of terms, those which describe the studied elements that represent and organize themselves in a hierarchical manner classes are extracted, considering a more general abstraction level towards specific classes (Rautenberg; Todesco; Gauthier 2009). In Figure 7, it is possible to see a sample of the services section.

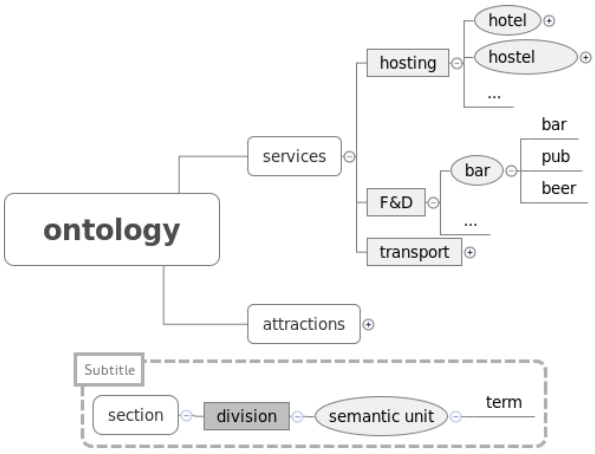

FIGURE 7: A SAMPLE OF APPLICATION ONTOLOGY

The four aspects that make up the elaborate application ontology include:

- Section: names that separate the elements investigated in this research: tourist services;
- Division: category used to analyze the results - similar to clustering;
- Semantic Unit: (SU) most specific significance level object within the study's assumptions;
- Terms: keyword(s) inserted on the system for a SU association were found and considered (s) as synonyms.

The application ontology was used to provide parameters for knowledge discovery among the collected data as well as assist in carrying out the search in an efficient manner. Figure 8 summarizes the phase 1.

I - Objectives, domains and other definitions	
Components	**Empirical Application in DMO's case**
Location	Facebook, Twitter and Youtube social medias.
Period	September 16th 2013 to August 31st 2014; divided by six periods to allow analysis over the time.
Ontology and terms	Tourism Attractions and Services in Curitiba application ontology.

FIGURE 8: THEORETICAL MODEL OF PROCESS AND EMPIRICAL ON DEFINING OBJECTIVES, DOMAINS AND SCOPES

Concerning the collection of data, as shown by Kalampokis, Tambouris and Tarabanis (2013), there are different approaches to this step ranging from manual and/or automatic selection techniques. Collecting and encoding a big amount of content on social media manually is tedious and time-consuming (He; Zha; Li, 2013). The present study adopted a dynamic approach that means there is no single and stable solution to data gathering. The search terms are obtained through various query searches in a computational process and are updated during the process based on an exploratory initial approach.

3.2. Technology tools and support definition

As Barbier and Liu (2011) pointed out, the problem itself determines the best approach and the tools required to solve it. Big data collection in social media is mainly supported by *Application Programming Interfaces* (APIs). These are resources that developers can use to create their own applications to collect public data from social media. Due to limitations on API interoperability between different social media, a third party solution, called Seekr Monitor, was selected to the data gathering of this research. This tool was tested and chosen based on the following characteristics that match the study goals: (i) connection with Facebook, Twitter and YouTube (initial target of this study); (ii) no limit on storage or messages collected based on estimated population of the WC2014; (iii) availability of data in an open interchange format, to proceed the framework path inside the domain of custom ICT tools; (iv) flexibility on strategies to obtain data to test and re-test the ontology application and its effectiveness.

One of the consequences of the lack of standards in API is the different understanding of a same question on search and find resources. In the following example (box), a specific target is defined: get all messages including a name of city (Curitiba) and its main attraction (Botanic Garden). In social media A, the Boolean search with both terms worked, but this was not the case in Social Media B. As a result, for each social media, a different strategy of data gathering had to be built when data came directly from API instances. Consequently, a third party software was developed to eliminate this barrier in the research.

Example:
Target: Botanic Garden of Curitiba: all the search results that include “Botanic Garden” and also “Curitiba”.
Boolean Search: “Botanic Garden” AND Curitiba
Result Social Media A: Correct, only results that include both terms.
Result Social Media B: Incorrect, results with the Garden, botanic and Curitiba appearing, but not necessarily together.

The process to select appropriate tools or solutions on data gathering depends on the targets set and settings. A test and re-test is also needed to legitimate the data from the source. Figure 9 summarizes the phase 2.

II - Technology tools and support definition	
Components	**Empirical Application in DMO's case**
Research	Looking for solutions in ICT that overcome API limitations and matches the settings of the study.
Test / Evaluation	Testing, analysis and evaluation of the software for a period of 15 days in a sample.
Definition	Definition of Seekr Monitor as data gathering tool.

FIGURE 9: THEORETICAL MODEL OF PROCEDURE AND PRACTICAL APPLICATION OF DEFINITION OF APPROPRIATE TOOLS AND TECHNOLOGY SUPPORT

3.3. Data Collection Strategies and Execution

Data collection in social media is a major challenge as it requires search for relevant information in an unstructured content environment. The work with knowledge discovery in databases, not just in a social media approach, usually is not a process that runs sequentially, but requires a repeated selection of samples and testing rounds to reach a best model (Kalampokis; Tambouris; Tarabanis, 2013). In this research, the challenge consists of four steps, illustrated in Figure 10 and explained below.

III - Data collection strategies and execution	
Components	**Empirical Application in DMO's case**
Collection Strategies	Tourist attractions in Curitiba; Tourism Services in Curitiba; hashtags (#) Curitiba and World Cup related.
Pre-tests	Floating reading of collected data samples.
Preprocessing	Preprocessing collection on file in CSV format.
Exportation	Collecting exportation in CSV format.

FIGURE 10: THEORETICAL MODEL OF PROCESS AND PRACTICAL APPLICATION OF DATA COLLECTION STRATEGIES AND EXECUTION

The first step is related to the collection strategies definition. A collection strategy begins through a query, a command that returns a result set according to the given criterion. Therefore, refers to setting the search terms to be monitored by the software, aiming to assist and perform the recovery of the content according to the characteristics defined previously. In short, queries are ‘questions’ that the monitoring software makes to the APIs to collect the contents of registered terms and, therefore, may require different approaches and strategies as shown in Figure 11.

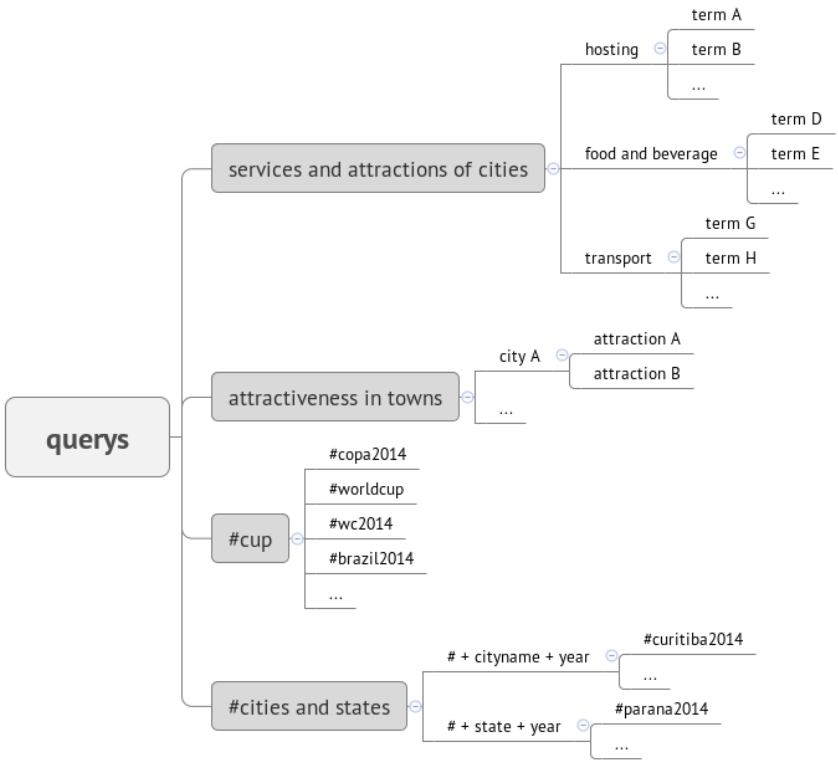

FIGURE 11: DATA COLLECTION STRATEGIES.

Four collection strategies were conducted in the first step. First and second strategies were based on respective ontologies of tourist services. On these approaches, the software recovered all cases where the city name appeared when at least one auxiliary term of the application ontology is with it. It is important to highlight that collections, rules, actions and search terms can be reconfigured, but they cannot be applied to previous messages collected, only in new messages. Therefore, the configuration of the searches and the rules are fundamental to the quality and relevance of the collected messages. Due to a common practice of social media users, complementary strategies using hashtags were created. Adding the hash symbol "#" to a term, the user associates his/her publication to a particular topic. These are assigned to photos, posts, videos and other content that allow users to find and bring together relevant content and conversations related to a particular topic (Kwak; Lee; Park; Moon, 2010). These third and fourth strategies were adopted as an alternative; considering the hashtags used in previous events by the official organization or by users. These hashtags can be direct references to the event - #cup section (i.e. #worldcup), or connected to the host cities - section #cities and states (i.e. #curitiba2014).

Second step consisted in the initial reading of the collected data, guided by the Bardin approach (2011) by establishing contact with the collected contents to get to know them by reading directly the results. When identifying inconsistencies in the collected data, the collection strategies may have to change in the *software* that is used for pre-processing of data - this is the third step. Fourth step consists of exporting and recording the collected content. This was done in this study by Comma Separated Values (CSV) format through the Seekr Monitor functionality. This step can also be attributed to the pre-processing of data due to the transformation of *noisy* data from social media into usable format. Such a format has been selected for being considered more functional by the amount of cases and interoperability.

3.4. Data Cleaning and Treatment

The need for treatment and cleaning of the data collected was evident from all SMM models studied. Cleaning and refining of data are efforts needed to reduce the major difficulties faced in SMM as the presence of irrelevant content (Chanana; Ginige; Murugesan, 2004; Barbier and Liu, 2011; Kim, 2013). In this study, the data cleaning and treatment processes were done by developing some tools and procedures. After exporting the collected contents in CSV file format, all the remaining data and information manipulation was developed internally through open source technologies. This option allowed more transparency and control. For this purpose, a UNIX server instance was created with MariaDB database, manipulated via script procedures in SQL (Structured Query Language), Python and PHP (Hypertext Preprocessor) languages. These technologies were employed to operationalize the later stages, the main scripts shown in Table 1.

TABLE 1: MAIN SCRIPTS DEVELOPED AND THEIR FUNCTIONS

SCRIPTS	FUNCTION
mysqlimport.py	Gather CSV files and import them to a preset MariaDB database table
expressoes.sql	Set the application ontology inside MariaDB database
mapaheuristico.py	Create compiled data files as input for the heuristic maps
pythonsql.py	Search the terms of the application ontology in stored messages and associate variables to the messages.
statments.sql	Control all of the data dealing process

Firstly, the results of four collection strategies were recorded into a single database so that all posts or cases in the survey sample would be treated together. Then, identical messages posted more than once by the same user were eliminated, forming only a database with univocal cases. Second step involved the treatment of inconsistent results, by development of new strategies to solve problems in the data. Based on CRISP-DM methodology, this step consisted also of the evaluation and review of the previously steps and to make sure that the process and the data are suited to the defined objectives. As with all quantitative and qualitative analysis, data quality is essential to obtain reliable results (Fayyad; Piatetsky-Shapiro; Smyth, 1996; Chapman et al., 2000; He; Zha Li, 2013; Abrahams et al., 2013).

Through a series of repetitions, it was possible to identify high-impact irrelevant groups of cases and treat it. This cycle consists of: (i) operating a pattern, which is associated with the existence - often - of occurrences that do not meet the goal; (ii) test and retest of dealings aimed at mapping the similar cases; (iii) execution of a final dealing with the purpose of excluding cases. Besides the presented dealings, other irrelevant cases were found such as ambiguity-affected cases. Even during the last stages, potentially inconsistent data may require that a new treatment cycle is made and executed. This feature is inherent to the research type, where the method itself is constantly reviewed and responses to the problem become more accurate. Figure 12 summarizes the phase 4.

IV - Data cleaning and treatment	
Components	**Empirical Application in DMO's case**
Results compilation	Results of data collection were stored in MariaDB running in a UNIX server instance, and manipulated with SQL and Python scripts.
Cleaning and treatment	Exclusion of cases longer than 500 characters; analysis of 100 most frequent messages from each division; identification and exclusion of inconsistent data.
Evaluation and revision of previous processes	Evaluation and revision of phases, stages and strategies for collecting, besides software and scripts settings update.

FIGURE 12: THEORETICAL MODEL OF PROCESS AND PRACTICAL APPLICATION OF DATA CLEANING AND TREATMENT

3.5. Data Mining

The data mining technique on this research was drawn from the models of Fayyad, Piatetsky-Shapiro and Smyth (1996), Han, Kamber and Pei (2012), CRISP-DM and methodologies proposed by Abrahams et al. (2013) and Hea, and Zha Li (2013). It consists of defining methods and data mining techniques to reveal unforeseen trends, correlations or patterns in the data, considering the general objective of the process.

On this study, the data was subjected to a systematic search for terms of application ontology. Aiming to associate cases with certain topics, similar to the study by Cvijikj and Michahelles (2011), weight was assigned based on the frequency of a term and number of mentions when similar terms were collected. This process is also analogous to the marking of categories recommended by Bardin (2011) and adapted to the study given the proportion of data. Figure 13 shows an example of the structure of the message object regarding the presence of analysis variables, already prepared for the consolidation of the results.

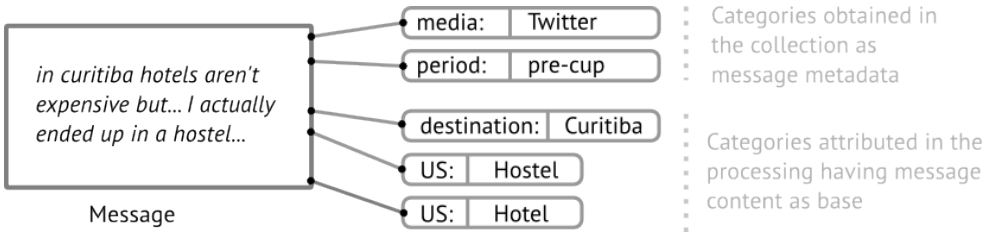

FIGURE 13: THE STRUCTURE OF THE MESSAGE OBJECT SCHEME READY FOR ANALYSIS

Before processing, each message already has a media and period variables associated with it, from the source. In the processing, firstly the target is identified through a search by name or synonymous terms. In the example, "Curitiba" is a term/pattern to designate the Curitiba destination and therefore a new link between message and category is made. In the next step, all terms associated with attractions section, city by city, are sought. In the example, no attraction was quoted thus no SU (semantic unit) link was made for this section. Then a query by terms of services section was performed, recovering, for instance, two SU - Hotel, Hostel. The same message can be associated with one or more SU's, on the understanding that both were cited and therefore have occurred. Even at that stage some cases can be excluded for not meeting the scope of the study, when the case does not present: media, time, destination and at least one SU.

V - Data mining	
Components	**Empirical Application in DMO's case**
Techniques definitions, settings and running	Automated systematic search for terms of ontology on database, associating cases with certain topics, semantic units, divisions and sections.

FIGURE 14: THEORETICAL MODEL OF PROCESS AND EMPIRICAL APPLICATION OF DATA MINING

A final round of message handling was also operationalized in this research. After separating only valid cases by the above criteria, an exploratory pre-analysis was made, based on the five most frequent messages for the destination.

3.6. Results Interpretation and Evaluation

This stage corresponds to post-processing, analysis, interpretation and evaluation of standards, issues, relationships and trends discovered and extracted by previous procedures. The criteria postulated in the study related to the presence or absence of elements of service ontologies as a way to measure popularity during periods of analysis. In this step the assignment of values across a set script in the database allowed to measure the amount of occurrences of the semantic units in each message collected in the database, the representativeness and the category of the ontology and the comparative studies between periods and other cities and tourist destinations. Also, the results and standards were associated and related to certain events, outdoor events and phenomena that due to user behavior were reflected in the database. Figure 15 summarizes the phase 6.

VI - Results interpretation and evaluation	
Components	**Empirical Application in DMO's case**
Post-processing	Values attribution through the script setup on the database to measure the amount of occurrences of the semantic units in each message collected in the database.
Analysis of results	Presence or absense of elements of the tourist attractions and tourism services ontologies to measure its popularity during defined analysis periods.
Interpretation	Results interpretation

FIGURE 15: THEORETICAL MODEL OF PROCESS AND PRACTICAL APPLICATION OF RESULTS INTERPRETATION AND EVALUATION

3.7. Results Visualization

Drawn from the model proposed by Han, Kamber and Pei (2012), this process refers to representing results using visualization techniques. Firstly, results were shown using tables by semantic unit (SU) and the average daily number of occurrences per period. In addition, an alternative way of data visualization was implemented to facilitate understanding and visualization of data and results. Therefore, the present study opted for representation with heuristic maps. The analysis categories, depending on the guiding ontologies present a series of belonging and frequency relations. For example, a tourist service belongs to a city, which belongs to a division - services. In addition, each service has a number of occurrences. The cross dependencies (relations) and scores (sizes) were expressed through circumscribed circles whose sizes represent the occurrences of density (Biz; Bettoni; Thomaz; Santos; Pavan, 2014). This solution, called heuristic map system, was developed at the Laboratory of Tourism, Technology, Information and Knowledge (Turitec, 2014), based on a data visualization

technology called D3 (Data-Driven Documents). This, together with a website developed internally, enabled the visualization according to the objectives of the research, with the possibility of filters for data sections and / or periods. A heuristic map was illustrated in Figure 16, by a screen shot of the developed system.

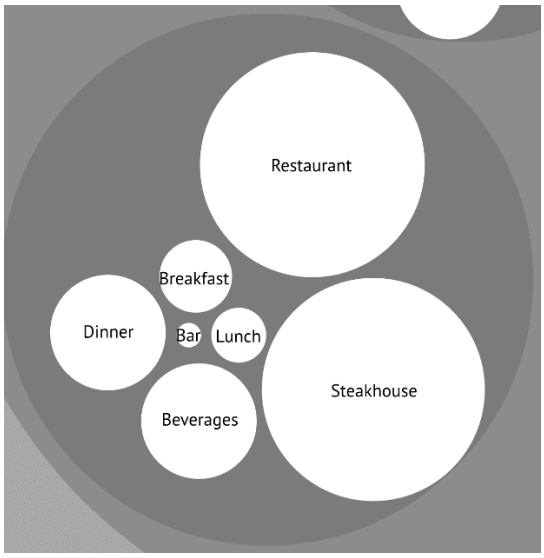

FIGURE 16: DISPLAY OF RESULTS BY MEANS OF HEURISTIC MAPS (TURITEC, 2014)

The current version of the system used shades as well as font sizes that have no meaning or relevance to the presentation of results. However, relations of belonging and size of the circles reflected the proportion of what was found in the survey. The screenshot on Figure 16 illustrates the SU from Curitiba tourist services present in the application ontology developed in this study. Figure 17 summarizes the phase 7.

VII - Results visualization	
Components	**Empirical Application in DMO's case**
Transformation and representation of results, patterns and possibilities	Tables and plots were made based on semantic unit (SU) and represent average amount of daily occurrences in each period of data collection; Heuristic Maps were also developed and used along Tourist Attractions and Tourism Services data from Curitiba.

FIGURE 17: THEORETICAL MODEL OF PROCESS AND PRACTICAL APPLICATION OF RESULTS VISUALIZATION

4. Results and Discussion

The monitoring data in the present study resulted from the practical application of the content mining framework presented in Figure 6 to the city of Curitiba, in terms of tourist services. This case study made it possible to develop a more accurate framework and to identify the application ontology components that had popularity fluctuations. Twitter results were validated and considered for the final analysis. This is related to difficulties in obtaining relevant data in other social media such as Facebook, Twitter and YouTube, due to limitations in its API and differences between the structures of content.

The results of the quantitative analysis are presented in tables and plots, by frequency and period. A total of 153,843 tweets were collected in the period, but 58,686 of these were considered valid (i.e. 38% of the collected content). The other 95,157 (i.e., 62% of the total) tweets were considered invalid because they were not assigned to any of the ontology terms attractions or they were invalidated for any of the databases treatment criteria. Importantly, the results presented support the claim by Paine (2011) that about 70% of the collected content is irrelevant. Table 2 shows the average number of events per day of services in Curitiba. The data was prepared using the criterion of popularity of the terms of ontology services in the periods defined in the survey. Green is the largest number of occurrences and red the least amount of occurrence being a maximum and minimum expression enhanced by the color and tone variation.

TABLE 2: NUMBER AVERAGE DAILY OCCURRENCE OF TOURIST SERVICES IN CURITIBA

	COLLECTION PERIODS					
US	**Pre-Raffle**	**Raffle**	**Pre-Cup A**	**Pre-Cup B**	**Cup**	**Post-Cup**
Hotel	0.0133	15,871	21,5517	24,3143	25,6286	8,2128
Hostel	0.0133	1,6774	0.931	1,2571	1,2571	0.766
Inn	0.0267	0.0645	0.069	0.0714	0	0
Accommodation	0	0.0484	0.0862	0.0286	0	0
Restaurant	0.1467	15,5161	17,1379	23,3571	10,2286	7,3191
Drink	0.0667	4,6935	3,1724	4,7714	4,5714	1,5957
Bar	0	1,6129	1,4828	1,1714	1,3429	1,0213
Lunch	0	3,3548	2,5517	0.7857	0.6571	2,3617
Breakfast	0.0133	0.9839	0.7586	0.2571	0.2857	1,0213
Dinner	0.04	2,8548	1,431	1,6571	0.9429	0.9574
Steakhouse	0.04	17,5	14,8793	15,3857	8,0571	4,3617
Airport	0.04	27,7581	46,8276	59,7714	39,4857	20,3404
Bus	0	0.0484	0	0.0286	0	0
Train	0	0	0	0	0	0
Car Rental	0	0	0	0	0	0
Taxi	0.0133	3,5968	2,0862	1,5429	1,0286	0.9787
Tour	0.0133	1,6452	0.3966	0.3429	0.4286	0.0213
Bicycle	0.0267	2,5806	3,8448	2,8286	1,6286	2,1277

The semantic units Airport, Hotel, Restaurant and Steakhouse stood out as Curitiba tourist services in Pre-Cup A in the analysis. Through manual content analysis of a sample about the services in Curitiba, it was possible to identify posts of users sharing compliments and some positive experiences on hotels, hostels, bars, restaurants, snack bars, public transportation, tours and activities, tourism products and services, as shown in Table 3.

TABLE 3: EXPERIENCES AND COMPLIMENTS ON PRODUCTS AND TOURIST SERVICES

CONTEXT	PUBLICATIONS
Hospitality	I was hosted here. Many nerd/geek in the environment and really good (@ Motter Home Curitiba Hostel) http://t.co/NhZw3lnezA
	Best place @ Motter Home Curitiba Hostel http://t.co/joyojE6aFd
Food and Beverage	*RT @danieltwa: First show in Brazil tonight in Curitiba tonight. So far Brazil's been great. I had chicken heart for lunch.*
	Big Holiday with the Family in Curitiba!!! Very good lunch in Santa Felicidade!!!!
	The best burger in the world! #madero #cwb #curitiba #brazil #beer #drinks #night #fritas #meat @… http://t.co/RycAZpQWFm
Transport	Here in Curitiba no metro, but public transportation works... organization is impressive …"
	I used Easy Taxi here in Curitiba and it`s fantastic!! Highly recommend!! Technology ar our service...
	Of the wonder that`s getting a cab in Curitiba in the rain and in two minutes. Thanks @easytaxi

Noteworthy publications was the bike tours offered by a local company and the Tourism Curitiba line, especially on the day that the bus collectors went on strike and tickets was not being charged. Some examples are shown in Table 4.

TABLE 4: PUBLICATIONS ON BICYCLE TOURS AND CURITIBA TOURISM LINE

CONTEXT	PUBLICATIONS
Bicycle Tours	Art *bike tour* through the street of #Curitiba with the @kuritbike gang. Lots of murals and curitiban graffitis. https://t.co/xy1CJBJGSS #streetart
	Sunny day in Curitiba, perfect for bike tour from art and cafés that I`m soon gonna make with the @kuritbike gang!!
	(Re)cognizing my own city in an art bike tour. #Curitiba @kuritbike http://t.co/2ZsKlghlop
Tourism Line	Great day sightseeing in Curitiba. The Oscar Niemayer museum is \"eye\" catching http://t.co/ENmbbczdnw
	Day of making Tourism Line in #Curitiba. First Stop Botanic Garden, with Wifi. Congratulations @Curitiba_PMC http://t.co/Oi2nMEXAcr
	Going for a ride in Curitiba. Tourism Line that goes through the main tourist spots of the city. It`s worth… http://t.co/1qxwCH7vEt
Tourism line and ticket collectors strike	Today is the day of making free tourism in Curitiba!! Bus from tourism line without Ticket collector! #grevedoscobradores
	Enjoying tourism line for free, hahahaha #Curitiba #LinhaTurismo #GreveBoa @ Tourism line http://t.co/qPu8Qf86oM
	Even the tourism line is for free today kkkkkk gone to know Curitiba kiddo lkkkkkkkkkk

However, despite the praise, it was also possible to identify critics especially about beer price in bars and restaurants. Also, the low number of restaurants that comply with the food safety legislation, the public transportation, and difficulty in getting accommodation during the FIFA World Cup 2014 period. Delay of works in the stadium was mentioned too. Table 5 presents some examples.

TABLE 5: CRITICS TO BARS, RESTAURANTS AND ACCOMMODATION

CONTEXT	PUBLICATIONS
Criticism for Bars and Restaurants	RT @_DANIELs: \"Curitiba is the most developed capital of the country\" a city where it`s so hard to get beer after 02:00am can`t be…
	Dear! I need to drink Stella Artois, but with the beer price in Curitiba it`s difficult! uahsuhuahsuhaushauhs
	RT @RobertaCanetti: Research shows that beer in Curitiba bars is more expensive than in São Paulo and Belo Horizonte - Band B. http://t.co/T…
Criticism of Accommodation	*Any Aussies going to the #worldcup had any luck with accommodation in Curitiba or Cuiaba? I'm striking out everywhere #socceroos;*
	@Tim_Vickery a basic hostel dorm bed is going for over £100 in rio and even £80 in Curitiba. Ridiculous! I was going to go but priced out
Criticism about the possibility to cancel the World Cup in Curitiba	*@hdrebner @BAHeraldcom FIFA needs to take a chill pill and consider those of us who've already booked accommodation and flights to Curitiba!*
	Sure hope the next news I hear about Curitiba is good. Got flights and accommodation riding on it. #worldcup #socceroos @FFA
	So those non-refundable flights and accommodation to Curitiba are looking pretty good right now. #WC2014

Main criticisms refered to the transport provision, with several mentions of the airport, bike paths, public transportation, taxis (service quality and local laws against mobile taxi apps), tourism and even the subway line (which does not exist, but was mentioned as a claim!), as shown in Table 6.

TABLE 6: PUBLICATIONS OF CRITICS ON TRANSPORT IN CURITIBA

CONTEXT	PUBLICATIONS
Airport	RT @hbrumjunior: @PrisciIlIa @FAXINANOPODER_ @mvdsister This is ADDITION OF INCOMPETENCE Curitiba Airport. http://t.co/XTf09N2qxh
	RT @jaimeendres: @CBNCuritiba you need to make a report about Curitiba airport. It's chaotic. RX queue. http://t.co/GpSdwoM9gO
	RT @summer_tanton: Curitiba Airport, yesterday. Mat in passengers' access to the plane and leak. #NaoVaiTerCopa @cbnbrasilia http://t.co…
Bicycle Tracks	@Curitiba_PMC hello PMC. Will there be isolation, with turtles or similar, from the bike track in this calm way of 7 de setembro?
	RT @gazetadopovo: Leia no blog Ir e Vir de Bike: Prefeitura investiu R$ 0,00 do orçamento nas ciclovias de Curitiba em 2013 http://t.co/obJ…
Public Transportation	"Passengers register supercrowding in bus station of Curitiba \http://t.co/AM3aLiP1oy"
	RT @Nossa_Curitiba: The way things are soon taxi will be cheaper than the bus in Curitiba
	RT @FILIPIOLIVEIRA: Curitiba today: no bus (drivers and ticket collectors on strike), no taxi (centrals don't answer), no patience (various co…
Taxi	"I'm defend taxi applications for the cell phone. In Curitiba radio-taxi/teletaxi have a deficiente service,... http://t.co/EPm4trobOW"
	Taxi drivers are fined and have their licenses seized for using EasyTaxi em Curitiba. Congratulations URBS, finding taxi in Curitiba is really easy!
	@Curitiba_PMC I hope so. Please be quick with this, we can't depend on radio taxi not in Curitiba not in SP.
Tourism Line	@Curitiba_PMC @lullylucky I visited Curitiba last weekend and I DON'T recommend this

CONTEXT	PUBLICATIONS
	tourism line. This sucks big deal!
	Reporter has to pay R$ 29 to make a report in the Tourism Line. We tried to help with the report take a look. @GustavoFruet @Curitiba_PMC
Curitiba Metro	"RT @espalhai_: Dilma announced for the 3rd time funding for the Curitiba Metro. Promised so much and did nothing. #PTMentindoNaTV"
	RT @manoelzoero: this Curitiba metro, same story since 2009
	RT @joaopaulom: @coroneldoblog BH beating Curitiba. Dilma announced metro here 6 times. But so far not even one centimeter... http://t…

Another consumer behavior trait that was evident through the content analysis was the photos and videos of dining experiences such as dishes and beers, frequented bars and restaurants, hotels or hostels where users stayed, among others. Table 7 shows some identified examples.

TABLE 7: PHOTOS AND VIDEOS POSTED

CONTEXT	PUBLICATIONS
Photos F&B	Almoço com os massas. #curitibacool #Curitiba #friends #brasil #almoço #instagram #instapic #instacool http://t.co/jG9DqdR1At
	#cerveja #beer #instabeer #ale #cerveza #brazil #brasil #curitiba http://t.co/d5UeHfyDeF
	Just published a photo @ Restaurante Madalosso - Santa Felicidade - Curitiba - PR http://t.co/7ErFJ6nPNo
Accommodation Photos	View from the hotel bedroom in Curitiba PR @ Curitiba Eco Hostel http://t.co/qG7BGsnmnk
	Just published a photo @ Curitiba Backpackers Hostel http://t.co/V67a0BZfBi
	Goooood moooooorning o/ #hostel #sacada #curitiba #pr @ Motter Home Curitiba Hostel http://t.co/gLIX2Trlve
Tourist activities photos	#curitiba #fmb3 #bike #bicicle #bici #bicicleta @ Museu Oscar Niemeyer (MON) http://t.co/tsrT2a8iOn
	#bikenight #bike #pedal #mon #curitiba @ Museu Oscar Niemeyer (MON) http://t.co/dJyMQDByu1
	bike day #curitibadebike #belleville #dialindo #soldaporra #verao #sendosaudavel #turistando #curitiba… http://t.co/1KOkKvTSHz
Videos	Curitiba Eco Hostel: http://t.co/GdFHPIOSV9 via @YouTube

The "check-in" functionality also demonstrated that food and beverage establishments were popular among users of social media. In Curitiba, the highlights were a restaurant, a steakhouse and three bars, according to examples illustrated on Table 8. Social Media users were often proud to announce were they were eating or drinking and often gave a positive public recommendation.

TABLE 8: CHECK-INS IN BARS, RESTAURANTS AND ACCOMMODATION

CONTEXT	PUBLICATIONS
Check-ins Steakhouses	I'm at Jardins Grill (Curitiba, PR) http://t.co/DlJJYjgOA5
	RT @ptg_r: I'm at Churrascaria Curitiba Grill II (Curitiba, PR) http://t.co/CZNyDNQXCC
	RT @LismaraContador: I'm at Alameda Grill (Curitiba, PR) http://t.co/4S8CjEZnjn

CONTEXT	PUBLICATIONS
Check-ins Restaurants	RT @_fer_santana: I'm at Madero Burger & Grill - @madero_bg (Curitiba, PR) http://t.co/TtIjA5yvkG
	RT @guilima78: I'm at Restaurante Spaghetto (Curitiba, PR) http://t.co/2r7lvXeW3g
	RT @BiaSobocinski: I'm at Restaurante Veneza (Curitiba, PR) w/ 4 others http://t.co/E66dAQ7NOl
Check-ins Bars	I'm at Barbarium Beer Pub (Curitiba, Paraná) http://t.co/5xS8dflwSy
	I'm at Mr. Beer (Curitiba, PR) http://t.co/8zQjk5xbIb
	I'm at A Varanda Beer House - @a_varanda_bar (Curitiba, PR) w/ 3 others http://t.co/ELGYt3l9Ti
Check-ins Accommodation	I'm at Motter Home Curitiba Hostel (Curitiba, PR) http://t.co/uvfwQuXKQM
	I'm at Curitiba Casa Hostel (Curitiba, PR) http://t.co/82u67LAnZG
	I'm at Curitiba Hostel (Curitiba, PR) http://t.co/veGaJy4XV0
	RT @FdeOD: I'm at Bourbon Curitiba Convention Hotel (Curitiba, PR) http://t.co/tZW5awUVVg

However, the bars may have highlighted the fact that the name has drafted terms of application ontology, and therefore may have influenced the results. To support this statement, it would be better to conduct a study and direct analysis and monitoring of the name of the establishments and, as appropriate, identify the mentioned establishments and the most popular among users. Another consumer behavior characteristic of tourism in social media that was evident was the search for tourist information and suggestions. In this study, it was possible to identify users requesting recommendations and tips on hotels and hostels in Curitiba to other users on social media, as examples shown in Table 9, supporting Customer to Customer cocreation.

TABLE 9: PUBLICATIONS OF QUESTIONS AND REQUESTS FOR ACCOMMODATION RECOMMENDATION

CONTEXT	PUBLICATIONS
Accommodation recommendations and tips	@patie_ do you know any cheap hostel/inn there in curitiba? I have a wedding there in the end of february!
	hostel or cheap motel by the road to spend the night in curitiba, does anyone know?
	I'm intending to go to Curitiba with Mozinho, where's the gang to indicate cool and cheap hotels or inns @curitiba
	Does anyone have a country house/inn/farm hotel near Curitiba to indicate??
Doubts	@Curitiba_PMC does the tourism line work normally morning?
	@Curitiba_PMC helpme! Tourism line is working fine, city hall?
	@Curitiba_PMC Mayor, my chayott! Do you know if during vacation the tourism line works on Mondays? Family would like to go for a ride tomorrow...

In addition to the users search on social media for information and recommendations, many users also published recommendations, news, travel hints and activities in the destination, answering to other users by their own free will. In addition to the conversations among users, public and private actors involved in the local tourism industry such as hotels, hostels, restaurants, and tour blogs, also participated actively publishing suggestions and tips on services, tourist attractions and activities to users and visitors of the city. Some examples in Table 10 illustrating the Customer to Customer experience cocreation

TABLE 10: RECOMMENDATIONS AND FOOD, DRINKS AND MEANS OF HOSTING TIPS

CONTEXT	PUBLICATIONS
F&B Recommendations and tips	RT @BlogCheckIn: The lunch time post to make everyone hungry. Meet the Madalosso, em Santa Felicidade, Curitiba: http://t…
	RT @Carlao_Picheth: Today Beer Fest in Bar Brahma Curitiba, international beers from Ambev such as Norteña, Hertog Jan Tripel, Patagônia Ambe…
	RT @FerRosse77: Check out these recommendations for dinner near Curitiba, PR. http://t.co/jGGoXT0Zb1 via @foursquare
Accommodation Recommendations and Tips	RT @TipTripViagens: Hostels of Curitiba testaded and approved http://t.co/6DbPuwGxWq via @TipTripViagens
	Check some inn options of the Metropolitan Region of Curitiba http://t.co/HAIUpbxARl

It was also identified that several publications aimed to promote and market destinations, tourism products and services, particularly profiles of hotels and *hostels*, as shown in Table 11. This is a form of indirect or direct promotion using social media engagement. In some occasions this was pure recommendation based on good experience and service received. In other occasions, it as direct advertisement, through engagement often aiming at Search Engine Optimization, through using key phrases.

TABLE 11: PUBLICATIONS OF PROMOTIONS AND MARKETING OF PRODUCTS AND TOURIST SERVICES

PUBLICATIONS
Curitiba Hostel has the best location of Curitiba, Historical Centerm – O Largo da Ordem, the favorite place for the tourists.
"Christmas atmosphere in Botanical Garden of Curitiba! Beautiful ain't it? Come here to know it! Stay at Hostel Roma! http://t.co/tsU64B6sLw"
The Receptive Tourism Center and bike tours in Curitiba are in Paraná TV report in the last... http://t.co/ewhgdTKyVj
RT @Curitiba_PMC: When visiting Curitiba, be sure to stroll the Tourism Line. This video shows how it is: http://t.co/49LCXXuCeu

Among the publications with promotional and business objectives, there was a highlight of the initiatives and marketing actions in social media developed by a local hostel. Table 12 shows examples of promotional messages and relationship published by Roma Hostel in Curitiba. This little accommodation property used twitter to amplify their message and attract guests.

TABLE 12: MARKETING AND USER ENGAGEMENT MESSAGES PUBLISHED BY ROMA HOSTEL CURITIBA

PUBLICATIONS
" The Roma Hostel Curitiba will give a Christmas present to our guests! \\ From 24 to 31/12, get 10% discount... http://t.co/20RNkSjJvq"
" The Curitiba Christmas schedule for today! \ Looking for Hosting? \ Stay here in Roma Hostel Curitiba!... http://t.co/JLrW3usxk4"
" Beautiful day in Curitiba! \ Here in Roma Hostel Curitiba you find this paradise within the city, to rest... http://t.co/xKmrUktYYn"
Where to stay in #Curitiba – #Hostel Roma http://t.co/JqQfBUPXN1 #brasil #ondeficar #sul #paraná

PUBLICATIONS
"Follow the Hostel Roma Curitiba at Instagram!\Follow us on Instagram!\@romahostel\#instagram\#romahostelcuritiba... http://t.co/zYyGYEvdzI"

5. Conclusions

This study provides a content mining framework and methodology in social media to support strategic and operational tourism management. The framework has been tested from the content mining on Twitter with the terms of the application ontology of tourist services in the city of Curitiba, Brazil. It proved to be effective on collecting and summarizing relevant content and identifying the terms of most popular touristic services components, as well as conducting quantitative and qualitative analysis. It reinforces that social media is an important source of information, communication and tourism promotion, requiring proper marketing strategies. At the same time, it is important to note that public and private local actors involved with tourism in Curitiba are using social media to achieve their goals.

Although SMM is an area already explored in previous studies, the dynamics of the content and the Internet make the models and techniques demanding constant evolution. One of the contribution of this study is the design and implementation of an empirical test combined with the formation of a framework in an area (i.e., tourism) that still lacks both theoretical studies and applied. This brings light to the tourist system of competitive advantage supported by SMM. The paper discusses and combines several methods and applications in SMM, text mining and content analysis to solve a problem in tourism, with the possibility of extension and replication in other areas. The framework provided offers a comprehensive, complete and more detailed model than in previous studies. The framework application provides an identification of user's compliments, opinions, positive and negative experiences about accommodation, restaurants, coffee shops, transportation, tours and activities among other tourism products and services. This also paves the way for other research fronts or even the evolution of a more general framework that addresses other areas. Perhaps once this model is operationalized and lessons have learned and incorporated in future versions, the model will support real time data collection, analysis and decision making to empower tourism dynamic product cocreation and value development. Public information shared on social media should enable analysis regardless of privacy legislation in different countries; although retargeting may require an opt in policy.

Aiming, at first, to test and validate the framework and its possibilities, the current study did not attempt a real-time monitoring. However, all technology support required and tools were developed since data gathering until data visualization. The effectiveness of real time monitoring depends on the subjects, number of social media posts, key words, data analyst and researchers as well as the tools and methods used. In the present study, due to the large number of publications, themes, keywords, technical and human resources of research, it was not possible to follow in real time. However, this model can eventually provide real-time implications for practitioners through an appropriate data analysis staff, analysis related to specific issues such as accommodation, transportation, safety, traffic, among other travel related issues. A real-time monitoring will allow DMOs to analyze these messages and use them strategically to redefine marketing and branding strategies. It will also support repositioning the city as a destination, anticipate problems related to infrastructure, assess the destination tourist attractions attractiveness, evaluate dynamically the tourist infrastructure and also analyze the

quality of tourist services providers to enhance tourist experiences and to strength destination brand and competitiveness. The study explains how tourism-related content posted by users on social media can be monitored and, consequently, be used as an information / knowledge source by the DMO and also private tourism companies to improve their competitiveness. This can enhance the strategic and operational management of tourism organizations and destinations.

For Curitiba as a destination on this test, the practical contribution of this study was to identify the popularity about tourist services, and understand who are the social media leaders from the community. It could also understand criticisms and critical problems reported by users and visitors in social media. Throughout all touristic related services, the Airport, Hotels and Hostels, Restaurants and the Steakhouse have the highest level of frequency, which is an opportunity to pursue a better understanding of the service at those positions. For instance, the Airport – as the most cited – has a delayed expansion plan on course during the FIFA World Cup 2014. Its popularity – positive or negative – could be explained partially by this fact. Eventually the methods will be developed to offer real time feedback and will challenge the destination authorities to manage their resources and those of stakeholders.

Although sentiment analysis was not performed on this study, an opportunity has been showed by the sample of messages. DMOs can choose a more specific and deeper (less wide) approach – extending the framework – and carry out targeted studies in the database about a particular topic of interest. For instance, selecting a few attractions (i.e. the most visited) and making an exploratory search inside those results, aiming to build a more specific ontology level (differing access, attractiveness, resources, points of interest, etc.). A sample overview, showed trends confirming some tourism consumer behavior characteristics in social media. For example, the use of hashtags (#), realization of "check-ins" in restaurants, hotels, bars, nightclubs among others, search and sharing tips, information, suggestions, reports and positive and negative traveling experiences demonstrate the importance of social media content for the competitiveness of tourism organizations.

The present study helps understand the importance, opportunities and benefits that the proposed process, monitoring and mining content in social media offers to tourism management and marketing. However, it is important to point out that monitoring and mining content in social media should not be summed up only in metric and collecting opinions, tourist experiences and user information. Qualitative evaluation of results can provide comprehensive overviews of destinations and in depth understanding of issues of strategic importance. Also, social media monitoring requires investment in training, human resources and new technologies so the DMO can collect data, interpret, analyze and study them. This activity can be transformed into knowledge, disseminated and used strategically in all governance and other public and private actors involved in the tourist activity.

Application of content mining in social media in tourism is still in the initial stage, and this was the main motivating factor for conducting this study. The framework is mainly composed of studies of other adjoining areas to the practical problem of the case in Tourism, which means it can be replicated and can present similar results, in other service areas and/or behavior consumers in social media. Problems already recognized as the high presence of irrelevant content, the need of customized technological solutions, the repetition of tests, and the maturity of an automatic classification system had to be overcome as steps in this framework. There are four limitations of this study. The first was the ontology of application, which have limited terms and needs to be expanded so that all possibilities and tourist services be included. As pointed out previously, effective solutions

require efforts and tests / re-tests of scope definition, strategies, ontologies and tools. The ontology coverage for tourism still needs to be developed. Semantic units like “inn”, “accommodation”, “bus” and “car rental” have no relevant results on the final dataset (Table 2), requiring a review of terms. Emerging categories / dimensions as “hosting tips” also needs to be added. The second limitation refers to data gathering and privacy as well as restrictions on social media to access public content of users, as this may have influenced the quantity and quality of collected cases. The third refers to the size of content collected, which made it impossible to carry out the manual content analysis for all messages. However, the analyzed sample was sufficient to identify the characteristics of the data. Finally, the fourth limiting factor was restriction to one social media for the final analysis. Nevertheless the SMM framework was developed and evaluated at this stage.

There are three suggestions for future research. The first relates to the development of a theoretical model of social media based knowledge management and content mining processes. This comprehensive framework should be implemented for DMOs or similar like governance bodies and other public and private stakeholders involved in tourism. The benefits and opportunities that the implementation of the content mining framework in this study will affect the competitiveness and development of tourism destinations. The second concerns the extension of the application ontology to new dimensions related to other public tourist organizations, private tourist organizations and future events. Future content mining research should concentrate on Sentiment Mining and Opinion Mining to collect data to understand the mood of tourism consumers in real time.

This research highlights the importance and richness of data and meaning that can be extracted from the content in social media. The main elements of the tourist trade in the opinion of social media can be made available through the implementation of the proposed framework. Likewise, the study verifies the challenge of dealing with content mining, even with a customized model for the problem. The choice of categorizing a priori by specialized ontology facilitates selection and data count, but it may also omit important information.The framework however makes a clear contribution to the development of SMM towards the creation a real time listening mechanism that will support tourism organizations and destinations to create dynamic processes that will enhance value added to all stakeholders in the ecosystem.

REFERENCES

ABRAHAMS, A. S.; JIAO, J.; FAN, W.; WANG, G. A.; ZHANG, Z. (2013). What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings. **Decision Support Systems**, 55(4), 871-882.

AMARO, Suzanne; DUARTE, Paulo; HENRIQUES, Carla. (2016). Travelers’ use of social media: A clustering approach. **Annals of Tourism Research**, 59, 1-15.

AMORIM, Sergio R. Leusin; CHERIAF, Malik. (2007). Sistema de indexação e recuperação de informação em Construção baseado em Ontologia. In: **III Encontro Tecnologia da Informação e Comunicação na Construção Civil-TIC2007**, Porto Alegre, Brasil.

BARBIER, Geoffrey; LIU, Huan. (2011). Data mining in social media. In: **Social Network Data Analytics**. Springer, 327-352.

BARDIN, L. (2011). **Análise de conteúdo**. São Paulo: Edições 70.

BIZ, A. A.; BETTONI, E. M.; THOMAZ, G. M.; SANTOS, C. K.; PAVAN C. S. (2014). **Relatório Técnico do Monitoramento de Mídias Sociais: Copa do Mundo FIFA 2014**.

BLUMRODT, J.; PALMER, A. (2013). Webpage Design and Quality of Seaside Tourism Destinations: A Question of Collaboration. **International Business Research**, 6(9), 1-13.

Boes, K., Buhalis, D., and Inversini, A., 2015, Conceptualising Smart Tourism Destination Dimensions, in Tussyadiah, I., and Inversini, A., (eds), **ENTER 2015 Proceedings**, Springer-Verlag, Wien, 391-404.

BUHALIS, Dimitrios. (2000). Marketing the competitive destination of the future. **Tourism Management**, 21, 97-116.

BUHALIS, D., and LAW, R., (2008), Progress in tourism management: Twenty years on and 10 years after the internet: The state of eTourism research, Tourism Management, 29(4), pp.609–623.

BUHALIS, Dimitrios; FOERSTE, Marie-Kristin. (2015). SoCoMo Marketing for Travel and Tourism: empowering co-creation of value. **Journal of Destination Marketing & Management**, 4(3), 151–161.

BUHALIS, Dimitrios; AMARANGGANA, Aditya. Smart tourism destinations. (2014). In: XIANG, Z.; TUSSYADIAH, I. (Eds.), **Information and communication technologies in tourism.** Dublin: Springer. 553–564.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. (2000). **CRISP-DM 1.0 - Step-by-step data mining guide**. SPSS Inc.

CHANANA, Vivek; GINIGE, Athula; MURUGESAN, San. (2004). Improving information retrieval effectiveness by assigning context to documents. In: **Proceedings of the 2004 international symposium on Information and communication technologies**. Trinity College Dublin, 86-91.

CHANDRASEKARAN, B.; JOSEPHSON, J. R.; BENJAMINS, V. R. (1999). What are ontologies, and why do we need them? **IEEE Intelligent systems**, 14(1), 20-26.

CHEN, Sherry Y.; LIU, Xiaohui. (2004). The contribution of data mining to information science. **Journal of Information Science**, 30(6), 550-558.

CHEN, Y.; AMIRI, H.; LI, Z. CHUA, T. S. (2013). Emerging topic detection for organizations from microblogs. In: **Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval**. ACM, 43-52.

CVIJIKJ, Irena Pletikosa; MICHAHELLES, Florian. Monitoring trends on facebook. (2011). In: **Proceedings of the 2011 IEEE Ninth International Conference**. 895-902.

CHUA, Alton YK; BANERJEE, Snehasish. (2013). Customer knowledge management via social media: the case of Starbucks. **Journal of Knowledge Management**, 17(2), 237-249.

CROOKS, A.; CROITORU, A.; STEFANIDIS, A.; RADZIKOWSKI, J. (2013). #Earthquake: Twitter as a distributed sensor system. **Transactions in GIS,** 17(1), 124-147.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. (1996). From data mining to knowledge discovery in databases. **AI magazine**, 17(3), 37.

FILIERI, R.; MCLEAY, F. (2014). E-WOM and accommodation: An analysis of the factors that influence travelers' adoption of information**. Journal of Travel Research**, 53(1), 44-57.

GRETZEL, Ulrike; YOO, Kyung Hyan. Use and impact of online travel reviews. (2008). In: **Information and communication technologies in tourism 2008**. Springer Vienna, 35-46.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. (2012). **Data mining**: concepts and techniques. Waltham: Elsevier.

HAYS, Stephanie; PAGE, Stephen John; BUHALIS, Dimitrios. (2013). Social media as a destination marketing tool: its use by national tourism organisations. **Current Issues in Tourism**, 16(3), 211-239.

HE, Wu; ZHA, Shenghua; LI, Ling. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. **International Journal of Information Management**, 33(3), 464-472.

KALAMPOKIS, Evangelos; TAMBOURIS, Efthimios; TARABANIS, Konstantinos. (2013). Understanding the Predictive Power of Social Media. **Internet Research**, 23(5), 544-559.

KAPLAN, A. M.; HAENLEIN, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. **Business Horizons**, 53(1), 59-68.

LEUNG, D.; LAW, R.; van HOOF H.; BUHALIS D. (2013): Social Media in Tourism and Hospitality: A Literature Review, **Journal of Travel & Tourism Marketing**, 30(1-2), 3-22

LIU, Z.; PARK, S. (2015). What makes a useful online review? Implication for travel product websites. **Tourism Management,** 47, 140-151.

MARINE-ROIG, Estela; ANTON CLAVÉ, Salvador. (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. **Journal of Destination Marketing & Management**, 4, 162-172.

MISTILIS, Nina; BUHALIS, Dimitrios; GRETZEL, Ulrike. (2014). Future eDestination Marketing Perspective of an Australian Tourism Stakeholder Network. **Journal of Travel Research**, 53(6), 778-790.

NEUHOFER, Barbara; DIMITRIOS, Buhalis; LADKIN, Adele. (2014). Co-creation Through Technology: Dimensions of Social Connectedness. In: ZHENG, X.; TUSSYADIAH, I. (Eds.), **Information and communication technologies in tourism 2014,** 339–352.

NEVES, Augusto José Waszczynskyj Antunes das; MARCHIORI, Patricia Zeni. (2014). Qualidade percebida em produtos e serviços em eventos: Técnicas e ferramentas para análise de conteúdo do Twitter. **Turismo & Desenvolvimento**, 2(21/22), 173-182.

PAINE, Katie Delahaye. (2011). **Measure what matters**: Online tools for understanding customers, social media, engagement, and key relationships. John Wiley & Sons.

PHILLIPS, P.; ZIGAN, K.; SILVA, M. M. S.; SCHEGG, R. (2015). The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis. **Tourism Management**, 50, 130-141.

RAUTENBERG, Sandro; TODESCO, José L.; GAUTHIER, Fernando A. O. (2009). Processo de desenvolvimento de ontologias: uma proposta e uma ferramenta. **Rev. Tecnol**, 30, 133-144.

SO, K. K. F.; KING, C.; SPARKS; B. A.; WANG, Y. (2016). The Role of Customer Engagement in Building Consumer Loyalty to Tourism Brands. **Journal of Travel Research**, 55(1), 64-78.

STATISTA. (2016). **Statistics and Market Data on Travel, Tourism & Hospitality**. Retrieved 26 April 2016, from http://www.statista.com/statistics/321509/frequency-of-travel-review-site-use-in-the-united-kingdom-uk/

TANG, Xuning; YANG, Christopher C. (2012). Social network integration and analysis using a generalization and probabilistic approach for privacy preservation. **Security Informatics**, 1(1), 1-14.

WILLIAMS, N.; INVERSINI, A.; BUHALIS, D.; FERDINAND, N. (2015). Community crosstalk: an exploratory analysis of destination and festival eWOM on Twitter. **Journal of Marketing Management**, 31(9-10), 1113-1140.

XIANG, Zheng; GRETZEL, Ulrike. (2010). Role of social media in online travel information search. **Tourism management**, 31(2), 179-188.

XIANG, Z.; MAGNINI, V. P.; FESENMAIER, D. R. (2015). Information technology and consumer behavior in travel and tourism: Insights from travel planning using the Internet. **Journal of Retailing and Consumer Services**, 22, 244-249.

XIANG, Z.; WANG, D.; O'LEARY, J. T.; FESENMAIER, D. R. (2015). Adapting to the Internet: Trends in Travelers' Use of the Web for Trip Planning. **Journal of Travel Research**, 54(4), 511-527.

ZAFARANI, Reza; ABBASI, Mohammad Ali; LIU, Huan. (2014). **Social Media Mining**: An Introduction. Cambridge University Press.

ZENG, B.; GERRITSEN, R. (2014). What do we know about social media in tourism? A review. **Tourism Management Perspectives**, 10, 27-36.