Tourism profiling: a semi-automatic classification model of points of interest

Amarildo Martins de MAGALHÃES
School of Information Science, Federal University of Minas Gerais
Belo Horizonte, Minas Gerais, Brazil

Renata Maria Abrantes BARACHO
School of Information Science, Federal University of Minas Gerais
Belo Horizonte, Minas Gerais, Brazil

Thomas MANDL
Information Science, University of Hildesheim,
Hildesheim, Germany

ABSTRACT[1]

Reviews are a powerful source of information that helps tourists in their decision-making process. However, using this volume of data to make decisions it is time consuming. For example, the city Foz do Iguaçu, located in Brazil, has more than 44k reviews on TripAdvisor. Based on these opinions, how could a tourist understand if this attraction is good for families, a romantic date, or if it offers a good outdoor experience? Moreover, which other attractions could offer similar experiences? These questions motivated this research, as we try to address the problem of classifying tourism attractions/destinations in profiles. We proposed a hybrid approach, using experts' knowledge and machine-learning with semi-automatic classification models to solve the problem. This paper presents a new approach to classify tourism attractions in profiles using reviews. Our findings show that, the most visited places are not necessarily the most relevant to a specific profile and as such the corresponding group of tourists. Understanding these profiles can aid the discovery or the selection of a travel destination. In addition, it allows governments and the private sector to target tourism marketing actions in the most assertive way.

Keywords: Information Classification, Machine-Learning, Reviews, Tourism Profiling.

1. INTRODUCTION

The paradigm of technological evolution has brought a disruptive change to people's behavior, as people now make decisions based on the content they consume on the Internet. Decisions, such as choosing a tourist destination and even more complex decisions, such as choosing a president, are made based on online content. More and more users can generate data and information and make it available through social media, allowing this content to spread easily. The methods used to classify, organize, and retrieve such information have become even more important because most of this content is unstructured [1]. In 2019, Tourism was responsible for 10.3% of the global Gross Domestic Product (GDP), and the industry created 330 million jobs, 10% of the world's total [2]. The Internet has been an important factor in this growth. In fact, the Internet makes it possible to offer tourist services in a different format that can reach more people in less time. Solutions such as Google, Airbnb, TripAdvisor and Booking offer services based on information which helps users in their decision-making processes [3]. Reviews of other people's experiences in a particular place have been an important source of information. Nielsen's survey [4] shows that 63% of all respondents mentioned that, before buying tourist products and services, they conduct some sort of online search. For example, the Foz do Iguaçu Waterfall attraction, located in the city Foz do Iguaçu (Brazil), has about 44k reviews on TripAdvisor. How could a tourist understand if this attraction is good for families, a romantic date, or if it offers a good outdoor experience? Are the most popular destinations also the most relevant for a specific group? Guy et al. [5] discovered that, due to the large volume of reviews available, users only read a few of them, thus losing important information. Understanding the profile of a tourist destination, involves the classification of its Points of Interest (POIs) which are attractions that can be experienced in a given destination [6]. However, identifying profiles in tourism is a complex task because it is related to technical and emotional factors [7].

The purpose of this research is to create a model to classify POIs in a tourist profile set defined by specialists. It presents a semi-automatic model based on the knowledge of tourism experts, and automatic text classification models that classify POIs based on millions of reviews. This investigation is based on linguistic analysis and considers the opinion of Brazilians who visited a particular

[1] We would like to express our gratefulness to Professor Thomas Mandl and Professor Allen Ronald DeSerranno for their detailed peer-editing of this document.

place. This approach explores a novel track in content-based tourism research. Moreover, to the best of our knowledge, this is the first attempt that uses reviews as a source to classify POIs in tourist profiles. The remainder of this article is organized as follows: section 2 discusses related work; section 3 describes data and the method used to classify POIs in tourist profiles; in section 4, we demonstrated the classification result followed by an exploratory analysis trying to answer research's questions; and finally, in section 5, we provided our conclusions by offering the research findings.

2. LITERATURE REVIEW

Automatic text classification

Automatic text classification generally involves: 1) pre-processing, 2) text representation, and 3) classification model. 1) Pre-processing accounts for the cleaning and preparation of the text to be represented. It is a natural language process responsible for removing items from the text that are not important for its meaning, such as stop-words (articles, prepositions, etc.), and techniques to organize the text, such as lemmatization and stemming or removing adjectives and verbs [12]. 2) Text representation is the transformation of text into numbers, using methods such as Bag of Words (BOW) or Term Frequency - Inverse Document Frequency (TF-IDF) [13]. 3) The next step in automatic text classification is the creation of a classification model. Machine Learning plays an important role here. Supervised Machine-learning models such as Logistic Regression, Random Forest, Support Vector Machine, and Bayesian Networks are used for automatic text classification [14]. The idea behind these classification methods is that they could learn from a labeled corpus (some text already classified into categories), and then predict the correct category of unseen text.

Tourism profiling

Tourist's characteristics influence their choices when searching for experiences, attractions, or tourist destinations. Tourism profiling is not a new theme, Cohen [15] showed that tourists' demands, needs, and expectations vary considerably according to their age, family environment and income range. Gibson and Yiannakis [16] presented tourist profiles from a pragmatic point of view with 17 roles, thus seeking to find characteristics that a particular tourist can assume, such as Escapist or Sun Lover. Their results suggested that psychological factors have a strong impact on the final decision. Some works tried to address the tourism psychological problem, using, e.g., the "big-Five" model, which, according to Soto [17], represents the 5 standard personality traits that every human being can have. To better understand this psychological aspect, Neidhardt et al. [19] proposed a new model for identifying tourist profile based on images and a seven-factor model, combined, reduced from 22 items (17 roles + 5 personality traits). This seven-factor model has the following profiles: *1) Sun and Chill-Out, 2) Knowledge and Travel, 3) Independence and History, 4) Culture and Indulgence, 5) Social and Sport, 6) Action and Fun, 7) Nature and Recreation*.

According to Sertkan, Neidhardt and Werthner [7], profiling and personalization techniques can aid the user's decision-making process regarding the selection of attractions and tourist destinations. The authors applied a classification method in a German tourism dataset. Although the algorithms demonstrated good performance, the authors concluded that a new study considering more data would describe the problem better and that one destination is linked to a set of tourism profiles, instead of just one. The authors have quoted the city Rio de Janeiro, which offers experiences for different profiles, such as *nature, nightlife, beaches* and others. This reinforces the idea of Gretzel et al. (2006), that people tend to choose destinations for a combination of characteristics instead of just one. Lawton and Kallai [20] pointed out that individuals recognize tourist attractions differently. Therefore, when classifying a destination or tourist attraction in profiles, it is important to observe a set of opinions. The literature highlights the importance of this theme and reveals some current limitations that we tried to address in this research.

Related work

According to Nielsen [21], 92% of people in the world say they trust more in recommendations of friends and family than any other form of marketing. In tourism, electronic word of mouth *(e-WOM*) is expressed in reviews written in portals such as TripAdvisor or Google about a restaurant, an attraction or a hotel. Reviews have been object of study of many researchers, either for their credibility, such as the work of Fang et al. [22], or for their ability to influence [23]. Recent research suggests that reviews on travel sites such as TripAdvisor, Booking and Yelp have increasingly influenced travel decision making [11]. Mckenzie and Adams [3] developed a destination similarity comparison based on reviews from TripAdvisor. The authors used a topic modeling (Latent Dirichlet

Information explosion somehow created a paradox: despite the large volume of data, Mutula [8] draws attention to the fact that only 0.5% of it is effectively analyzed and used. The Internet is the most used channel when it comes to searching for information and, considering Big Data, information retrieval needs to evolve to effectively transform raw data into actionable knowledge [9]. When users consume online content, they are forming judgments based on credibility, usefulness, accuracy, or bias [10]. These judgments influence their decisions [11]. This influence is affected by how easily the user can use and interpret the content, and the volume and unstructured nature of the text can become a barrier in this process. Automatic text classification has been object of study to overcome this problem.

allocation – LDA) algorithm to verify similarities of destinations and identify differences in opinions of tourists from different countries. Our study differs from their approach, as it tries to classify tourist attractions using a profile set, rather than focus on similarities between cities.

Guy et al. [5] tried to solve the problem of the massive amount of review data by extracting tips based on POI reviews. The authors used external experts to validate the results and 73% of the tips extracted were classified as useful. Arentze, Kemperman, and Aksenov [6] developed a model for a recommendation system for generating custom travel itineraries based on POI features. The authors validated their model through an online questionnaire and the results revealed substantial differences in demands and needs among tourists. Sertkan, Neidhardt and Werthner [7] use the seven-factor model to automatically capture similarities of destinations to use in a recommendation system. Although similar to our approach, rather than reviews, the authors use a relational database to map destinations into the seven-factor model. The work of Shin et al. [24] is similar to our approach. However, instead of tourist profiles, the authors used the concept of Destination Personality Scale (DPS).

3. DATA & METHODS

This work offers a new perspective using TripAdvisor reviews as a source for tourism profiling. We used a hybrid approach relying on expert's knowledge and supervised machine learnings methods.

Domain Expert knowledge elicitation

According to Cleverley and Burnett [25], the synergy of using mixed methods (manual and automatic) can lead to better results than a single approach in the knowledge organization process. The expertise of two agents with more than 10 years of experience in the tourism domain contributed to our research. Based on the literature and in the expert's knowledge, we created a profile set. Each profile could include psychological factors and interests for specific group. To understand a profile characteristic, experts chose the 3 most relevant and similar POIs for each group. Table 1 shows the profile set created:

Table 1 – Profiles and the 3 most relevant POIs

Profile	Profile items	3 most relevant POIs
Culture	Culture, knowledge, history	Rome-Roman Forum, Paris-Louvre Museum, Buenos Aires–Bellas Artes National Museum
Landscape / Architecture	Hills, modern or old buildings	Madrid-Temple of Debod, Machu Picchu, Prague–Prague Castle
Nightlife	Casinos, bars, nightclubs	Las Vegas-Casino at Bellagio, Porto Seguro-Mucugê Street, Lisbon-Bairro Alto
Family	Fun with kids, parks	Orlando – Walt Disney World Resort, Arraial d'ajuda – Eco Parque, Orlando – Universal Orlando Resort
Gastronomy	Wines, beers and cuisine	Bento Gonçalves-Casa Valduga Winery, Petrópolis–Bohemia Beer House, Belo Horizonte-Maletta Building
Adventure	Diving, hiking	Canela-Alpen Park, Natal-Genipabu Dunes, Las Vegas-Stratosphere Tower
Beach	Beaches	San Andres - San Luiz Beach, Maceió - Ponta Verde Beach, Florianópolis – Ingleses Beach
Shopping	Malls, fairs and stores	Orlando- International Premium Outlets, Miami-Disney Springs, Miami-Dolphin Mall
Relax	Rest, reflection, escape	Paris-Seine River, Veneza-Grand Canal, Jericoacoara-Pôr do Sol Dune
Romantic	Places to enjoy for two	Gramado–Negro Lake, Punta cana- Saona Island, Buenos Aires-Puerto Madero
Nature / Exotic	Rivers, waterfalls, animals, and exotic options	Bonito-Lago Azul Cave, São Paulo-Botanical Garden, Dubai-Burj Khalifa
Religious	Churches, cathedrals and religious history	São Paulo–Guarapiranga Sacred Soil, Roma-Basilica di Santa Maria Maggiore, Paris-Notre-Dame cathedral

The most relevant POIs for each profile included Brazilian and international cities.

Data

The data was accessed through the TripAdvisor portal with the strict object of study and non-commercial use. For each tourist destination, TripAdvisor can display information about hotels, restaurants, flights, attractions and other information to help tourists make decisions. A particular tourist destination can be classified according to the attractions it has [3]. These POIs can include museums, parks, monuments, malls, castles, churches or beaches [4]. Each destination has its POIs and each POI has its reviews. Reviews from 2009 to 2019 were retrieved from the POIs of 148 destinations (63 Brazilian cities). We choose the most visited cities by Brazilians with the experts help. The following fields were retrieved: *review title, comment, date of visit, attraction,* and *date of review*. We extracted only reviews written in Portuguese language by users who indicated that they lived in Brazil. We set an exclusion criterion: POIs with less than 30 reviews and destinations with less than 30 POIs would not be used, to avoid influence of outliers during the classification. After this selection, 124 destinations, 2,627 attractions and 3,427,594 reviews were available for model building and analysis in the research. Based on the experts' knowledge, we considered the positive reviews (rating >= 4 out of 5) of the three most relevant POIs as a labeled corpus.

Classification model

A semi-automatic model to classify POIs in profiles was developed. The model is semi-automatic because it uses the experts' knowledge to create a labeled corpus. Figure 1 shows an example of Culture profile:

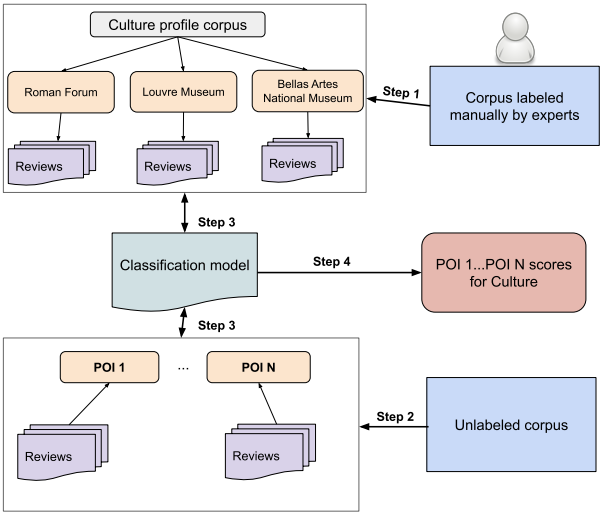

Figure 1 – POIs Classification Strategy for Culture Profile

The strategy was to create a labeled corpus for each profile using the reviews from the 3 most relevant POIs as mapped by experts (Step 1). For each profile, we used the 3000 most rated and recent reviews. This normalization is important, as a different number of reviews could bias the result. We also created an unlabeled corpus using all reviews for each POI (Step 2). Based on the profile labeled corpus, the classification model would learn how to identify the similarity between tourism profile and the POI's unlabeled corpus (Step 3). Further, the classification model finds the similarity score for each POI and profile (Step 4).

We followed the traditional three steps: *preprocessing, text representation, classification models*. For both corpus, we applied common preprocessing methods using the Natural Language Toolkit (NLTK) for Portuguese [26]. A review is the concatenation of review title and comment. We used terms in the singular form in lowercase, removing accents, numbers, special characters, and punctuation from each text. Bigrams and trigrams were applied for terms that appeared 5 times together. One situation we had to deal was the NLTK limitations in the Portuguese language. We improved the NLTK stop-words list, adding 1,323 new terms. We also created a Portuguese list with 260,524 verbs and 6,626 adjectives. This is a technical contribution to this work, considering the current limitation of tools for Portuguese natural language processing (NLP) tasks. We removed from each corpus stop-words, verbs and adjectives using the lists created. We applied the TF-IDF method to transform each corpus into numbers. In TF-IDF, each review becomes a vector and we applied the Euclidean norm (l2-norm) to normalize the vectors length. Furthermore, we created 4 classification models using the Sklearn implementation of the supervised algorithms Random Forest, Support Vector Machine, Multinomial Naïve Bayes and Logistic Regression [27]. Results are shown in the following section.

4. RESULTS

Classification model evaluation

The cross-validation approach was applied to verify the classification models' performance. We split the corpus of 36,000 labeled reviews into 25% for test and 75% for learning. The test samples are randomly selected keeping almost the same quantity of reviews for each profile. The test corpus has 9,000 reviews and the learning corpus has 27,000 reviews. We removed the profile information from the test corpus, so, the algorithms should predict the correct profile for each review in the test corpus. Comparing the predicted value with the correct value allows us to measure information retrieval metrics such as *Precision, Recall* and *F1-Measure.* We used the k-fold (k=5) test, which allows us to test the performance of each algorithm in a sample subgroup of reviews in the test corpus. Table 2 shows the performance of each algorithm:

Table 2 – Models performance results

Model	Precision	Recall	F1 - Score	Accuracy
Support Vector Machine (SVM)	0.7300	0.7288	0.7270	0.7314
Logistic Regression	0.7392	0.7254	0.7238	0.7289
Naive Bayes	0.7171	0.7163	0.7135	0.7178
Random Forest	0.5799	0.5277	0.5246	0.5317

The SVM model presented a slightly better accuracy. We tested different algorithm parameters to find out the best scenario. Following this result, we used a confusion matrix to analyze the results considering the profiles. The confusion matrix shows the correct profile of each review and the predicted one with a sum per (predicted x correct). The *Relax* profile had the best results by SVM with 803 reviews correctly classified out of 994 tested, which represents a performance of 89.83%. Nightlife had the worst performance with 643 reviews being classified correctly out of 981, which represents a performance of 65.54%. This profile was confused 81 times with the Landscape/Architecture profile.

Classification of tourism attractions – POIs

The similarity between POI and profile was computed using the classification models applied in the POI's unlabeled reviews. In total, 2,627 POIs were classified in each profile for all supervised models (SVM, Random Forest, Naïve Bayes and Logistic Regression). The Table 3 shows the example of the Louvre Museum POI profile classification using SVM:

Table 3 – Louvre Museum profile classification

Profile	SVM Score	Normalized
Culture	0.29756	0.40350
Landscape / Architecture	0.09474	0.12650
Religious	0.08191	0.10890
Nature/Exotic	0.07056	0.09340
Romantic	0.06261	0.08260
Relax	0.06106	0.08040
Shopping	0.06093	0.08030
Family	0.05814	0.07650
Night life	0.05731	0.07530
Gastronomy	0.05600	0.07350
Adventure	0.05598	0.07350
Beach	0.04319	0.05600

The Culture profile is the most similar and relevant for the Louvre Museum attraction with a similarity of 0.29756, or considering the normalized value, 0.40350 between 0 and 1. The second most similar profile is Landscape/Architecture with a normalized value of 0.12650. The Louvre offers paintings, art, sculptures, and different types of collections for different audiences of the cultural profile. On the other hand, the imposing architecture of the building that houses the collections also draws attention, as does the glass pyramid at the entrance.

Tourism profile rankings

Since we could classify each POI in the 12 profiles, it was possible to explore profile tourism information such as: which Brazilian city is the most similar to Paris? What are the top 10 most romantic attractions or destinations? To exemplify the rankings, we show the ten most relevant destinations for the Nightlife profile in Table 4.

Table 4 – Ten most relevant destinations for Nightlife profile

Destination	Relevance	Best scored POI
Las Vegas	2.1293	Casino at Bellagio
Lisboa	2.0021	Bairro Alto
Arraial d'Ajuda	1.8246	Mucugê St
Montreal	1.7771	Rue St-Paul
Cartagena	1.7764	Bairro Getsemani
Montevidéu	1.7615	Sofitel Montevideo Casino
Punta del Este	1.7447	Conrad Casino
Madrid	1.6937	La Latina
Barcelona	1.6873	El Born
México City	1.6292	La Condesa

The first three destinations are exactly those chosen by specialists within the most characteristic POIs of the Nightlife profile. Among the 5 most relevant attractions in Las Vegas, two are casinos (Bellagio and Wynn) and two are streets with casinos (The Strip and Fremont Street). Only one Brazilian destination appeared in the top 10, the city Arraial d'Ajuda. Using the classified corpus, it is also possible to find similar destinations or attractions, comparing their value in each profile using probability distribution measures such as Euclidian distance. The most similar city compared to the Brazilian city Ouro Preto is Tiradentes, another historic city in the state Minas Gerais. However, the second most similar city is Brussels in Belgium. Las Vegas's most similar city is Dubai, which is also known as an exotic city surrounded by a desert environment.

Relation of popularity and profile relevance

This category seeks to answer the question: are the most popular destinations on TripAdvisor the most relevant for a given profile? To answer this question, the experts helped to map the profiles in TripAdvisor categories. This allows them to check the popularity of a group of categories (profiles) adding up the number of visits each destination received in some period. For the analysis, we used the 20 destinations with the highest number of visits in each profile, that is, the 20 most popular in the TripAdvisor categories that are mapped to that profile. To exemplify this result we show the comparison of the Culture profile, because it was the profile with the highest

number of related attractions in TripAdvisor. The list of most popular destinations for Culture includes national and international destinations. Figure 2 shows the comparison Scatter graph with the two dimensions studied (Popularity x Relevance):

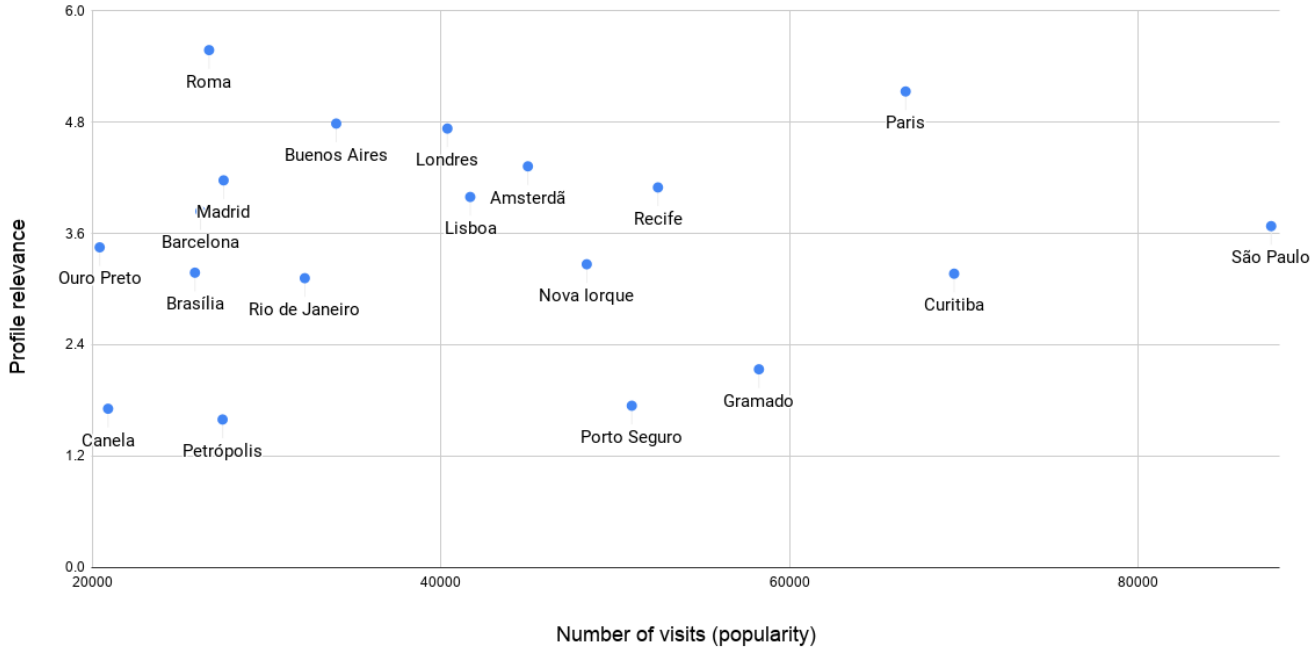

Figure 2 – Destination profile relevance x destination popularity
Source: Research Data

In Figure 2, São Paulo is shown as the most visited destination according to TripAdvisor, with more than 80 thousand visits between 2013 and June 2019. Although being the most popular destination, São Paulo is not the most relevant destination if we consider Culture, being only in the 10th position with 3.68 points. The most similar destination to the Culture profile was the city of Rome in Italy with a score of 5.57 points. However, if we look at the popularity scale, Rome appears only in 15th position, with about 26 thousand visits in the period studied. Paris appears as the most popular and most relevant city for Culture, ranking 2nd in popularity with around 66 thousand visits, and with a Culture relevance score of 5.13 points. Considering only Brazilian destinations, Recife is the most relevant destination for Culture. The result of the Culture profile shows that, although there are destinations among the most popular and relevant ones, such as Paris for example, there are also interesting and relevant destinations for Culture, but with less popularity. The same behavior was observed for all the other profiles. This factor may be related to the content-based approach using reviews where variables such as education and socioeconomic level could impact the results.

5. CONCLUSIONS

Reviews are an important source for information about a product, a service, or a tourism attraction, however, for one user, it is time consuming to read all the available content. Online reviews are the new WOM, and consumers make decisions using them. As information production and sharing grows exponentially, new methods to summarize and analyze the amount of data are necessary. We presented, in this research, a hybrid approach to cope with this situation in the tourism domain. Using the expert's knowledge, we created a semi-automatic model to classify attractions and destinations in tourism profiles. The integration between expert's efforts and technological methods, such as provided by machine learning concepts, could drive the retrieved information and organization process to better results. Moreover, preprocessing is essential for the results quality. As presented here in our study, we removed verbs and adjectives improving NLP datasets for Portuguese language. This factor can influence the classification quality as nouns are more specific to the attraction's characteristics, and verbs and adjectives are more generic. Our results show that it is possible to classify tourism attractions or destinations considering reviews and using a semi-automatic approach. The accuracy of 0.73 achieved using the SVM algorithm can be considered as very good because, in our approach, we choose to classify the POI positive reviews rather than classifying the review itself.

Our findings show that although being the most visited, some destinations are not the most relevant to some specific public (profile). São Paulo is the most visited Brazilian city in Culture profile according to TripAdvisor, however, Recife, another Brazilian city, presents more Culture relevance. Also, a tourist could select one similar destination instead of others, due to factors such as distance, budget, or weather, such as the example of Ouro Preto and Brussels. This is important, as it becomes

possible to offer new destination options to a specific group of tourists. Understanding the weaknesses and strengths of POIs in each profile allows governments and the private sector to target tourism marketing actions at a specific public, by creating or removing new related attractions, and so on. Tourists, on the other hand, could get similar experiences based on the tourist profiles they obtain from a recommendation system. A tourist could get a POI or destination recommendation based on his profile, as a recommend movie on Netflix for example. The information available in reviews is a key factor to tourist decision-making and organizing and grouping this user-contributed content in profiles, allows a new perspective in deciding where to go or what to do. Finally, it seems fair to observe that profiles also open an opportunity to explain algorithmic recommendations to users and therefore, contribute to explainable AI.

6. ACKNOWLEDGEMENTS

This work was cooperatively supported and financed by Federal Institute of Education, Science and Technology of Minas Gerais, and the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES).

7. REFERENCES

[1] S.K. Paul, M. Agrawal, S. Rajput, S. Kumar, "An Information Retrieval (IR) Techniques for text Mining on web for Unstructured data", **International Journal of Advanced Research in Computer Science and Software Engineering**, Vol. 4, No. 2, pp. 67-70.

[2] World Travel & Tourism Council (WTTC), **Travel & Tourism Economic Impact 2019 World**, London: WTTC, 2019.

[3] G. McKenzie, B. Adams, "A data-driven approach to exploring similarities of tourist attractions through online reviews", **Journal of Location Based Services**, Vol. 18, No. 2, 2018, pp. 94–118. doi: 10.1080/17489725.2018.1493548

[4] Nielsen Company, **Global connected commerce survey**, http://www.nielsen.com/us/en/insights/reports/2016/global-connected-commerce.html, 2016.

[5] I. Guy, A. Nus, A. Mejer, F. Raiber, Extracting and ranking travel tips from user-generated reviews. In **Proceedings of the 26th International Conference on World Wide Web**, pp. 987–996. International World Wide Web Conferences Steering Committee, 2017. doi: 10.1145/3038912.3052632

[6] T. Arentze, A. Kemperman, P. Aksenov, "Estimating a latent-class user model for travel recommender systems", **Information Technology & Tourism**, Vol. 19, 2017, pp. 61–82. doi: 10.1007/s40558-018-0105-z

[7] M. Sertkan, J. Neidhardt, H. Werthner, "What is the "Personality" of a tourism destination?", **Information Technology & Tourism**, Vol. 21, 2018, pp. 105–133. doi: 10.1007/s40558-018-0135-6

[8] S. Mutula, "Big Data Industry: Implication for the Library and Information Sciences", **African Journal of Library, Archives and Information Science**, Vol. 26, No. 2, 2016, pp. 93-96.

[9] B. Hjorland, "Knowledge Organization (KO)", **Knowledge Organization**, Vol. 43, No. 7, 2016, pp. 475–484.

[10] S. Sen, D. Lerman, "Why are you telling me this? An examination into negative consumer reviews on the web", **Journal of Interactive Marketing**, Vol. 21, No. 4, 2007, pp. 76–94.

[11] B.A. Sparks, H. Perkins, R. Buckley, "Online travel reviews as persuasive communication: the effects of content type, source, and certification logos on consumer behavior", **Tourism Management**, Vol. 39, 2013, pp. 1–9. doi: 10.1016/j.tourman.2013.03.007

[12] Among the 5 most relevant R. Baeza-Yates, B. Ribeiro-Neto, **Modern Information Retrieval**, Essex: ACM press, 1999.

[13] S. Qaiser, R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents", **International Journal of Computer Applications**, Vol. 181, No. 1, 2018, pp. 25–29.

[14] P. Louridas, C. Ebert, "Machine Learning", **IEEE Software**, 2016, pp. 110-115.

[15] E. Cohen, "Towards a sociology of international tourism", **Sociological Research**, Vo. 39, No. 1, 1972, pp. 164–182.

[16] H. Gibson, A. Yiannakis, "Tourist roles needs and the lifecourse", **Annals of Tourism Research**, Vol. 29, No. 2, 2002, pp. 358–383. doi: 10.1016/S0160-7383(01)00037-8

[17] C.J. Soto, Big five personality traits. In M.H. Bornstein, M.E. Arterberry, K.L. Fingerman, J.E. Lansford (eds), **The SAGE encyclopedia of lifespan human development**, Thousand Oaks: Sage, 2018. pp. 240–241.

[18] M. Braunhofer, M. Elahi, F. Ricci, User personality and the new user problem in a context-aware point of interest recommender system. In I. Tussyadiah, A. Inversini (eds), **Information and Communication Technologies in Tourism 2015**, Cham: Springer, 2015. pp 537–549. doi: 10.1007/978-3-319-14343-9_39

[19] J. Neidhardt, R. Schuster, L. Seyfang, H. Werthner, Eliciting the users' unknown preferences. In **Proceedings of the 8th ACM Conference on Recommender systems**, pp. 309-312. Foster City: ACM, 2014. doi: 10.1145/2645710.2645767

[20] C.A. Lawton, J. Kallai, "Gender differences in wayfinding strategies and anxiety about wayfinding: a cross-cultural comparison", **Sex Roles**, Vol. 47, No. 9/10, 2002, pp. 389–401.

[21] Nielsen Company, **Consumer trust in online, social and mobile advertising grows**, https://www.nielsen.com/us/en/insights/article/2012/consumer-trust-in-online-social-and-mobile-advertising-grows/, 2012.

[22] B. Fang, Q. Ye, D. Kucukusta, R. Law, "Analysis of the perceived value of online tourism reviews: influence of readability and reviewer characteristics",

Tourism Management, Vol. 52, 2016, pp. 498–506. doi: 10.1016/j.tourman.2015.07.018

[23] S. Shin, N. Chung, D. Kang, C. Koo, How far, how near psychological distance matters in online travel reviews: a test of construal-level theory. In A. Inversini, R. Schegg (eds), **Information and communication technologies in tourism 2016**, Cham: Springer, 2016. pp. 355–368. doi: 10.1007/978-3-319-28231-2_26

[24] S. H. Shin, S. B. Yang, K. Nam, C. Koo, "Conceptual foundations of a landmark personality scale based on a destination personality scale: Text mining of online reviews", **Information Systems Frontiers**, Vol. 19, 2017, pp. 743–752. doi: 10.1007/s10796-016-9725-z

[25] P. H. Cleverley, S. Burnett, "The Best of Both Worlds: Highlighting the Synergies of Combining Manual and Automatic Knowledge Organization Methods to Improve Information Search and Discovery", **Knowledge Organization**, Vol. 42, No. 6, 2015, pp. 428–444.

[26] S. Bird, E. Loper, E. Klein, **Natural Language Processing with Python**. O'Reilly Media Inc, 2009. https://www.nltk.org, 2020.

[27] Scikit-learn: **Machine Learning in Python**, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. https://scikit-learn.org/stable/supervised_learning.html#supervised-learning, 2020.