Proceedings of the Federated Conference on Computer Science and Information Systems pp. 81–84
DOI: 10.15439/2020F179
ISSN 2300-5963 ACSIS, Vol. 21

An extensive analysis of online restaurant reviews: a case study of the Amazonian Culinary Tourism

Luiz Carlos Fernandes Junior*, Jorge Silva Junior*, Antonio Jacob Junior† and Fábio Lobato*†
*Federal University of Western Pará, Santarém, Brazil
†State University of Maranhão, São Luís, Brazil
Email: {luizcarlossfjr, jorgeluizfigueira, antonio.jacob}@gmail.com, fabio.lobato@ufopa.edu.br

***Abstract*—Analyzing User-Generated Content present in social media has become mandatory for companies looking for maintaining competitiveness. These data contain information such as consumer opinions, and recommendations that are seen as rich sources of information for the development of decision support systems. When observing the state of the art, it was found that there is a lack of antecedents that address the analysis of online reviews of Brazilian restaurants. In this sense, the focus of this work is to fill this gap through a case study of Santarém city. The results show that professionals in this segment can use these analyzes in order to improve the user's experiences and increase their profits.**

I. Introduction

IN 2018, the tourism sector contributed US$ 152.5 billion to the Brazilian Gross Domestic Product (GDP)[1]. In the city of Santarém (Pará, Brazil), located in the very heart of the Amazon rainforest, the collaboration of this sector is significant. According to the municipal secretary of tourism [1], this activity injects about US$ 32 million in the local economy, driving segments like restaurants, hotels, travel agencies, bars etc.

The internet has completely changed the way the information related to tourism are distributed and consumed [2]. The User-Generated Content (UGC) growth has a significant impact on the tourism sector, influencing travelers in the decision-making process [3]. According to [4], UGC is all forms of content created, disseminated, and consumed by users.

Restaurant reviews are useful for the known segment as culinary tourism. In summary, this kind of tourism enables the recognition of values related to a certain territory's culture, so gastronomy is transformed into tourist products. In this panorama, [5] points out that online restaurant reviews influence consumers' decision-making, which is vital to the companies' analysis of this information to improve their services [6]. In the last years, with the data volume available on the internet and diversity growth, many challenges regarding data collection and analysis in this sector have emerged [7], [8]. One of these is to analyze the immense volume of textual data, a task practically impossible to be performed manually [9]. To tackle this obstacle, computational techniques such as Text Mining can be employed in order to identify patterns and generate insights that can support the decision making process [10], [11], [6].

Through a literature review, it was realized a lack of antecedents that explore the knowledge extraction from UGC on social media in Brazilian restaurants. In addition, the related works do not address the correlation of the authors' gender with relevant topics considered by them. In this context, the present work aims to analyze patterns extracted from restaurant reviews on the TripAdvisor platform, carrying out a case study of the city of Santarém. For this purpose, Text Mining techniques were applied to answer the following Research Questions (RQ):

- **RQ-1** What is the predominant sentiment expressed in the TripAdvisor reviews of restaurants in the Santarém, and which genre of customers has the most negative reviews?
- **RQ-2** What are the patterns of positive and negative comments?
- **RQ-3** What are the main topics covered in TripAdvisor reviews of restaurants in the Santarém - Is there a distinction of themes between the male and female gender?
- **RQ-4** How do the identified topics relate to each other?

The remainder of this paper is structured as follows. In Section II the related works are presented. The Experimental Framework used is described in Section III. The results and insights are discussed in Section IV. The conclusions and future works directions are given in Section V.

II. Related Work

Analysis in UGC has been widely addressed in several application domains due to the potential in the process of improving services and products [12]. This information is even more important for the hospitality sector, whose audience considers it to be a very reliable method of decision-making [13], [3]. In this scenario, a large part of this information is made up of textual data, so an appropriate approach to analyze this massive data is the use of text mining techniques [6].

Among the use of these techniques, are highlighted the Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) for the topic modeling task, as described in the studies by [14], [15]. By analyzing data collected from TripAdvisor, Airbnb, Couchsurfing, and Booking platforms, the authors obtained the segmentation of the type of review (for instance: *comfort, location, and experience*) as well as the identification of main service problems (for instance: *levels of cleaning service*).

[1]Data from Brazil's Ministry of Tourism

In [16], the authors used a classifier based on Naïve Bayes and in [17] the python libraries NLTK and TextBlob were used to identify the polarity of reviews collected from the TripAdvisor and Yelp platform. As a result, both papers present the distribution of consumers' opinions regarding the service. In [16], they combine topic modeling with sentiment analysis to obtain the main topics segmented by polarity. Besides, in [17] used the python library Guess-Gender to identify the gender of the authors' reviews in order to determine differences in assessment methods according to gender.

III. Experimental Framework

In this Section, the experimental framework used in this study will be described.

A. Data Acquisition

The TripAdvisor [2] platform was used as a data source, given its prominent position in the tourism field. All comments from Santarém restaurants that had at least one review on the site by April 2020 were collected, summing up 3,881 reviews from 186 restaurants. The data extracted include: i) restaurant name, ii) restaurant score, iii) review title, iv) comment score, v) review content and vi) username. A web crawler written in Python was developed to extract the data and the informations extracted were stored in a file in the Comma-Separated Values (CSV) format.

B. Data pre-processing

The pre-processing step was divided into two parts. The first one performs the filtering of assessments based on readability. Using the Flesch Kincaid index adapted to Portuguese[18], the most readable comments (scores between 75 and 100) were selected. Given that these are more influential in other customers' decision-making[19], the analyzes on this new dataset tend to reflect more accurate results for business managers.

The second one, consists of handling the information to remove inconsistencies and improve the results reliability [20]. This process was conducted using the NLTK library because it has support for the Portuguese. The following steps were performed: (i) characters conversion to lower case; (ii) accentuation substitution; (iii) punctuation and special characters removal; (iv) numbers deletion; (v) stopwords removal; (vi) emojis elimination.

C. Sentiment Analysis

Sentiment analysis can be defined as a technique for handling opinions, feelings and subjectivity in texts [21]. The Polyglot [22] library was used to perform this task, as it had good results in previous works for Portuguese language [23]. Since this library produces a numerical result (P) that varies from -1 to 1, the values obtained in this analysis were categorized as Neutral when $P = 0$; Positive when $0 < P \leq 1$ and Negative when $-1 \leq P < 0$.

[2]https://tripadvisor.com/

D. Topic Modeling

This task was divided into two stages: i) topics extraction; ii) the correlation analysis of the topics identified.

Regarding the first stage, the Non-Negative Matrix Factorization (NMF) technique was used, given its efficiency for text mining tasks in short documents [24]. The weight used to represent the values of these words in the term matrix was the Term Frequency-Inverse Document Frequency (TF-IDF) because good results were achieved with it in previous text mining tasks [25]. After performing the extraction, it is necessary to generate a results annotation, this step being performed manually from a subjective analysis by the authors [26].

To find the best coherence between the number of topics and the number of words, an analysis of the coherence of this relationship was performed using the metric Pointwise Mutual Information (PMI). Regarding the second stage, the topics correlation was conducted to verify which topics are most related to each other. Each topic represents a set of terms and each term is associated with one or more comments. Thus, the following elements were considered: the nodes represent the topics; the edges represent the relationship between the topics, and the greater the thickness of the edge, the more intense the correlation between the topics.

IV. Results

Initially 3,881 reviews were obtained, and from these, the readability analysis resulted in of 794 reviews helpfulness to conduct the other analyzes. From an analysis of the database, it was noted trend for users to give a high rating score, so that the lowest scores (10 and 20) together have only 38 occurrences.

After the pre-processing step, the sentiment analysis algorithm was applied. There were 62.5% positive comments, 22.3% neutral and 15.2% negative. Given this scenario and the first part of the **RQ-1**, it is possible to conclude that the main polarity expressed is positive. Examples of positive and negative comments can be seen, respectively, in items 1 and 2, 3 in Table I.

TABLE I: Examples of pre-processed comments.

Item	Comment after pre-processing stage
1	has wonderful view location amazing food served great price is worth
2	price pasties expensive size quantity filling are offered quality ingredients good problem cost x benefit pastel meat has wind can catch cold care meat cheese served satiate hunger great
3	location waterfront santarem think unique self service kilo city serves barbecue prepared food rolls prawns bad price simple environment kinda tight food more

The gender identification was performed manually and of a total of 794 usernames, 357 (45.0%) were recognized as being male, 312 (39.2%) female and 125 (15.8%) undefined. The comments classified as undefined are justified by the presence of pseudonyms, such for example, *Dream508624* and *Y4979PGalinem*, being therefore impossible to label them.

In this panorama, considering only the comments in which it was possible to identify the author as belonging to one of the

genders, and correlating these data with the sentiment analysis results, the Figure 1 is presented. Thus, it is possible to answer the second part of the **RQ-1**, in which the male gender obtained a slightly higher occurrence of negative comments. So, there are no significant differences between gender in the negative comments.

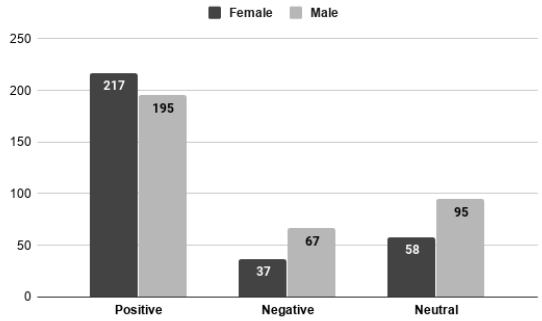

Fig. 1: Correlation between sentiment analysis and gender.

In order to understand how the pattern of a positive or negative comment is characterized, the rule extraction was conducted considering the sentiment polarity as labels. Similar to the analysis conducted in [25], the Decision Tree algorithm was used and, as data entry a BoW representation of the comments with the binary weight scheme. The result obtained was a set of descriptive rules presented in Table II.

TABLE II: Rules extracted *per* polarity.

Class	Rules	Coverage
Positive	Absence of: eat, delay, fried, hunger, leave, high, rotten, waiting, stairs	82,66%
Negative	Occurrence of: good, food, variety. beach, road, liked,	16,52%

When analyzing the extracted rules, it is possible to answer the second research question (**RQ-2**), in which the comments whose sentiment is considered positive tend to have the absence of terms such as "*delay*", "*fried*", "*hunger*", "*leave*" and "*rotten*". From these terms, it is possible to infer that the majority of customers take into consideration the waiting time and food quality. Regarding the negative comments, the occurrence of terms such as "*good*", "*food*", "*variety*", "*beach*", and "*road*" does mention the menu quality, location, and accessibility.

The the topics coherence analysis was conducted based on three viewpoints: i) all 794 comments; ii) only female comments; and, iii) only male comments. The best found for this viewpoints are respectively: combination of the 5 topics of greatest coherence with 100 topics and their 20 main words; combination of the 5 most coherent topics with 80 topics and their 10 main words; combination of the 5 most coherent topics with 100 topics and their top 5 words.

Analyzing Table III, it is possible to answer the first part of the **RQ-3**, in which it is noted that the main topics present in the comments are customer service, location, menu, space/infrastructure and appetizers.

TABLE III: Most prevalent topics.

Mode	Topics	Coherence average
All dataset	Menu, Space/Infrastructure, Customer service, Location, Appetizers	4.38
Female reviews	Environment, Menu, Location, Infrastructure, Customer Service	4.12
Male reviews	Location, Menu, Pricing, Options/Varieties, Customer Service	4.38

Analyzing also the Table III, it is possible to notice some distinct topics addressed between male and female reviews. For example, the topic pricing and options/varieties only appears in males, while infrastructure and environment were more consistent in females. Thus, it is possible to answer the second part of **RQ-3**: there are differences between the aspects addressed by people of different genders in online reviews.

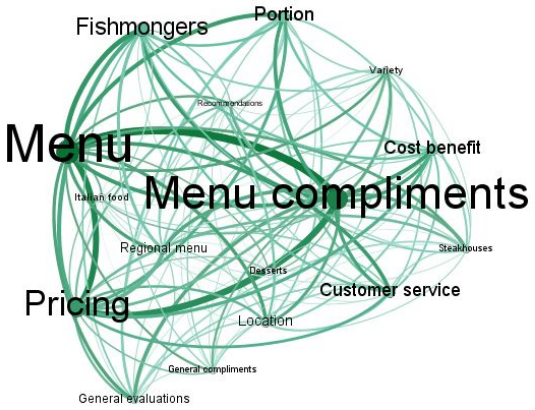

Fig. 2: Relationship between reviews topics.

The correlation between the topics can be verified through the analysis of the Figure 2. The size of the words represents the weight of the node in the network. The variation in tones and the thickness of the edges represent the intensity of the relationship. In this context and considering the **RQ-4**, is possible to highlight that comments that compliment the restaurant's menu usually contain evaluations referring to the price, showing the strong relationship between these attributes. In addition, the consumption of fish by tourists is evident, a fact justified by being a characteristic dish of the region.

The results obtained have practical implications:

- Most of consumer's are using the platform to give positive feedback or only to describe the restaurant facilities and services;
- Service, location, menu, space and appetizers are the most relevant aspects reported in the restaurants reviews;
- Based on identifying the most relevant topics by gender, restaurant managers could offer gender group discounts,

considering that these have different specific needs.

V. Conclusion

In this work, four tasks related to text mining were performed in order to extract relevant knowledge from online restaurant reviews: sentiment analysis, identification of the author's gender, extraction of rules, and topic modeling. In a brief contact with practitioners working directly with restaurants sector, we could validate the knowledge extracted. In this context, we conclude that the UGC is a rich source for the extraction of relevant knowledge from online restaurant reviews, taking the author's gender as a basis.

Our results can contribute to the management of companies related to culinary tourism, helping them to develop better products and services, centered on consumer expectations. Our work has some limitations, for instance, the emojis were removed in the pre-processing phase, which can impact on the sentiment analysis results. In future work, we would like to resolve these limitations, delve further into the modes of assessment by gender and extend the scope of the research, considering other locations and business domains.

Acknowledgment

The authors would like to thank the Maranhão Foundation for the Protection of Research and Scientific (FAPEMA) and the Tutorial Education Program (PET-IEG-UFOPA) for supporting the development of this work.

References

[1] G1, "Turismo em Santarém cresce em 2018 e injeta R$ 176 milhões na economia, aponta estudo," https://g1.globo.com/pa/santarem-regiao/noticia/2019/02/11/turismo-em-santarem-cresce-em-2018-e-injeta-r-176-milhoes-na-economia-aponta-estudo.ghtml. Accessed 21 April 2020., 2019.

[2] J. Navío-Marco, L. M. Ruiz-Gómez, and C. Sevilla-Sevilla, "Progress in information technology and tourism management: 30 years on and 20 years after the internet-revisiting buhalis & law's landmark study about etourism," *Tourism Management*, vol. 69, 2018. doi: https://doi.org/10.1016/j.tourman.2018.06.002

[3] Y. Narangajavana Kaosiri, L. J. Callarisa Fiol, M. A. Moliner Tena, R. M. Rodriguez Artola, and J. Sanchez Garcia, "User-generated content sources in social media: A new approach to explore tourist satisfaction," *Journal of Travel Research*, vol. 58, no. 2, 2019. doi: https://doi.org/10.1177/0047287517746014

[4] A. J. Kim and K. K. Johnson, "Power of consumers using social media: Examining the influences of brand-related user-generated content on facebook," *Computers in Human Behavior*, vol. 58, 2016. doi: https://doi.org/10.1016/j.chb.2015.12.047

[5] S. Lee, H. Ro *et al.*, "The impact of online reviews on attitude changes: the differential effects of review attributes and consumer knowledge." *International Journal of Hospitality Management*, vol. 56, 2016. doi: https://doi.org/10.1016/j.ijhm.2016.04.004

[6] S. Schmunk, W. Höpken, M. Fuchs, and M. Lexhagen, "Sentiment analysis: Extracting decision-relevant knowledge from ugc," in *Information and Communication Technologies in Tourism 2014*. Springer, 2013, doi: https://doi.org/10.1007/978-3-319-03973-2_19.

[7] B. G. Nistoreanu, L. Nicodim, and D. M. Diaconescu, "Gastronomic tourism-stages and evolution," in *Proceedings of the International Conference on Business Excellence*, vol. 12, no. 1. Sciendo, 2018. doi: https://doi.org/10.2478/picbe-2018-0063

[8] G. J. Miller, "Comparative analysis of big data analytics and bi projects," in *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2018. doi: http://dx.doi.org/10.15439/2018F125

[9] A. Klein, M. Riekert, and V. Dinev, "Accurate retrieval of corporate reputation from online media using machine learning," in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2019. doi: http://dx.doi.org/10.15439/2019F169

[10] R. Talib, M. K. Hanif, S. Ayesha, and F. Fatima, "Text mining: techniques, applications and issues," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 11, 2016. doi: https://doi.org/10.14569/IJACSA.2016.071153

[11] Y. Zhao, *R and Data Mining: Examples and Case Studies*, 12 2012. ISBN 978-0-12-396963-7

[12] F. Lobato, M. Pinheiro, A. Jacob, O. Reinhold, and Á. Santana, "Social crm: Biggest challenges to make it work in the real world," in *International Conference on Business Information Systems*. Springer, 2016. doi: https://doi.org/10.1007/978-3-319-52464-1_20

[13] L. Yan, N. Cha, H. Cho, and J. Hwang, "Video diffusion in user-generated content website: An empirical analysis of bilibili," in *2019 21st International Conference on Advanced Communication Technology (ICACT)*. IEEE, 2019. doi: https://doi.org/10.23919/ICACT.2019.8701897

[14] C. Marcolin, J. L. Becker, F. Wild, G. Schiavi, and A. Behr, "Business analytics in tourism: Uncovering knowledge from crowds," *BAR-Brazilian Administration Review*, vol. 16, no. 2, 2019. doi: https://doi.org/10.5748/9788599693148-15CONTECSI/PS-5707

[15] G. Santos, M. Santos, V. F. Mota, F. Benevenuto, and T. H. Silva, "Neutral or negative? sentiment evaluation in reviews of hosting services," in *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, 2018. doi: https://doi.org/10.1145/3243082.3243091

[16] V. Taecharungroj and B. Mathayomchan, "Analysing tripadvisor reviews of tourist attractions in phuket, thailand," *Tourism Management*, vol. 75, 2019. doi: https://doi.org/10.1016/j.tourman.2019.06.020

[17] M. P. Silveira, W. Z. Xavier, and H. T. Marques-Neto, "Análises de dados de sistemas crowdsourcing: estudo de caso de avaliações de estabelecimentos realizadas no yelp," in *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*. SBC, 2018. doi: https://doi.org/10.5753/brasnam.2018.3593

[18] T. B. F. Martins, C. M. Ghiraldelo, M. d. G. V. Nunes, and O. N. de Oliveira Junior, "Readability formulas applied to textbooks in brazilian portuguese," 1996.

[19] B. Fang, Q. Ye, D. Kucukusta, and R. Law, "Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics," *Tourism Management*, vol. 52, 2016. doi: https://doi.org/10.1016/j.tourman.2015.07.018

[20] D. Cirqueira, M. F. Pinheiro, A. Jacob, F. Lobato, and Á. Santana, "A literature review in preprocessing for sentiment analysis for brazilian portuguese social media," in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2018. doi: https://doi.org/10.1109/WI.2018.00008

[21] N. Rodríguez-Barroso, A. R. Moya, J. A. Fernández, E. Romero, E. Martínez-Cámara, and F. Herrera, "Deep learning hyper-parameter tuning for sentiment analysis in twitter based on evolutionary algorithms," in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2019. doi: http://dx.doi.org/10.15439/2019F183

[22] Y. Chen and S. Skiena, "Building sentiment lexicons for all major languages," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 383–389.

[23] L. Rodrigues, A. Junior, and F. Lobato, "Disability-Related News: An Analysis of User-Generated Content on Social Media Posts," in *In Proceedings of the 16th National Meeting on Artificial and Computational Intelligence*. SBC, 2020. doi: https://doi.org/10.5753/eniac.2019.9336

[24] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, "Experimental explorations on short text topic mining between lda and nmf based schemes," *Knowledge-Based Systems*, vol. 163, 2019. doi: https://doi.org/10.1016/j.knosys.2018.08.011

[25] J. Silva Junior, R. Rossi, and F. Lobato, "A Lyric-Based Approach for Brazilian Music Knowledge Discovery: Brazilian Country Music as a Case Study," in *In Proceedings of the 16th National Meeting on Artificial and Computational Intelligence*. SBC, 2020. doi: https://doi.org/10.5753/eniac.2019.9348

[26] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Leveraging multi-domain prior knowledge in topic models," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.