Tourism-Related Placeness Feature Extraction From Social Media Data Using Machine Learning Models

P. Muñoz[1], E. Doñaque[2], A. Larrañaga[3], J. Martinez[4]*, A. Mejías[5]

[1] Department of Financial and Accounts Economics, University of Vigo. Pontevedra (Spain)
[2] Possible Incorporated S. L.
[3] CINTECX, University of Vigo. Pontevedra (Spain)
[4] Department of Applied Mathematics, University of Vigo. Pontevedra (Spain)
[5] Department of Business Organization and Marketing, University of Vigo. Pontevedra (Spain)

Received 1 February 2022 | Accepted 2 December 2022 | Early Access 21 December 2022

Abstract

The study of *placeness* has been focus for researchers trying to understand the impact of locations on their surroundings and tourism, the loss of it by globalization and modernization and its effect on tourism, or the characterization of the activities that take place in them. Identifying places that have a high level of placeness can become very valuable when studying social trends and mobility in relation to the space in which the study takes place. Moreover, places can be enriched with dimensions such as the demographics of the individuals visiting such places and the activities the carry in them thanks to social media and modern machine learning and data mining methods. Such information can prove to be useful in fields such as urban planning or tourism as a base for analysis and decision-making or the discovery of new social hotspots or sites rich in cultural heritage. This manuscript will focus on the methodology to obtain such information, for which data from Instagram is used to feed a set of classification models that will mine demographics from the users based on graphic and textual data from their profiles, gain insight on what they were doing in each of their posts and try to classify that information into any of the categories discovered in this article. The goal of this methodology is to obtain, from social media data, characteristics of visitors to locations as a discovery tool for the tourism industry.

Keywords

M3 Inference, Machine Learning, Social Media, Tourism, Word2Vec

DOI: 10.9781/ijimai.2022.12.003

I. Introduction

Tourism is a source of wealth and sustained growth. The relationship between tourim and economic growth has been explained in several studies [1], [2]. Until 2019, the main motivation for international travel was tourism, being the reason for 56% of them, followed by visiting friends and family, health, religion and other purposes (27%), and business travel (13%). Until that time, tourism was the world's third largest export, after chemicals and fuels [3]. International tourism spending experienced an average annual growth of 4.6% between 2010 and 2018 [4], 2020). Even in 2019, this sector grew by 3.5%, contributing 10.3% to global GDP (i.e., Gross Domestic Product) and 28.3% to global exports of services [5]. Tourism has also undergone a qualitative transformation. The externalities associated with mass tourism, the emergence of a great diversity of tourism products to face competition from destinations, the change in the profile of the tourist who seeks different experiences focused on culture, nature, authenticity, among others [6], [7] have been some of the factors associated with this transformation.Tourism activities related to the cultural heritage-nature binomial the backbone of cultural tourism [8].

* Corresponding author.
E-mail address: javmartinez@uvigo.es

The concept of cultural heritage (i.e., CH) is very broad, including elements such as landscapes, historical sites, works of art, biodiversity, traditions, social values, sensory experiences, among others. In its contemporary meaning, it is made up of a tangible or physical component, the tangible cultural heritage; and an intangible component, the non-material cultural heritage [9]–[11]. The valuation of CH has been imposed in the narrative that guides the design of tourism products, due to the great dynamizing potential of the society and economy of the territories where it is promoted [12]. On the other hand, it has favored the development of cultural heritage tourism, as a tourism with a global dimension [13].

However, a tourist location may succeed to the point of generating a problem, of discomfort of residents [14] or disturbance of the environment [15]. Furthermore, the location of the tourist destination is dynamic [16]. In the context of globalization and increased human mobility, placeness can be affected by geopolitical issues, wars, terrorism, security threats and health emergencies among others [17], [18], [19] .

In this sense, the study of the factors that influence the formation of placeness of tourism products is increasingly being investigated [20], [21]. *Placeness* is defined as the uniqueness of a place determined by the set of its natural and historical features, its tangible and intangible cultural assets, emotions and sensations that the place can generate both to the inhabitants of the destination and tourists [22],[23]. Social networks such as Instagram are a valuable source of data for the study of the aforementioned factors that influence the creation of *placeness*.

Please cite this article in press as:
P. Muñoz, E. Doñaque, A. Larrañaga, J. Martinez, A. Mejías. Tourism-Related Placeness Feature Extraction From Social Media Data Using Machine Learning Models, International Journal of Interactive Multimedia and Artificial Intelligence, (2022), http://dx.doi.org/10.9781/ijimai.2022.12.003

Thus, drawing on previous work that defined the concept, the availability of social media data and machine learning and data mining techniques, we intend study the viability of extracting *placeness* features in accordance to the ontological approach proposed by [24]. We expose here a case study that focuses on the demographic and activity information in a clearly defined area and period of time. We intend to show-case a small scale study with the goal of introducing the methods as a first stage, but we aim to extend this to larger studies covering several locations in a wide area as a mean to discover new potential touristic opportunities.

A. Related Works

Previous research groups have used social media to infer information about people or places and have developed tools and methodologies which provide promising results. Noe, the two key articles on which our work is based are described [25], [26], [27], [28].

- Inferring *placeness* from Starbucks In [24] the authors focus their research around *placeness* as "the sense of place" and its importance in architecture or urban design. In their study they define an ontology for placeness which describes it as a relation to four factors: place, visitors, time and activity. They came up with a novel methodology to extract *placeness* features from Instagram posts and characterize a specific location given the information inferred for the given factors. They show-cased an study conducted from posts tagged in Starbucks in three major cities and compared its results to the current big data based approach and showed promising results from the relatively small amount of data they dealt with.
- Inferring *demographics* from social media A deep learning system is specified in [29] as an alternative to infer demographic data from social media users while providing resilience against biases that favour dominant languages and groups. The results achieved in their study were very promising, reaching accuracies higher than other demographic inference systems while providing better support for text in different languages, supporting up to 32 languages.

II. Methodology and Mathematical Background

Drawing on the works presented, a particular inference pipeline was set up to enrich Instagram posts with the same dimensions defined in [24] with the support of the M3 Inference pre-trained model to deal with cases where the images do not offer demographic information based on face recognition [30] [31]. On top of that, an education estimator was added based on a linguistic formula to measure the readability score of posts [32].

A. M3 Inference

According to Z. Wang et. al [29], this deep learning system is named after its multimodal, multilanguage and multi-attribute inference capabilities because it has been designed and trained integrating different machine learning models to extract features closer to the input, to support 32 languages and infer three different attributes. Fig. 1 shows the input data the model requires and the output is obtained from it.

This system's architecture combines DenseNet [33] [34] for image classification, and a 2-stack bidirectional character-level Long Short-Term Memories [35] neural networks as input models that are later combined in what they call a *dropout layer* followed by, in sequential order, two densely connected layers, one Rectified Linear Unit activation layer and three different output layers that apply a SoftMax [36] function to the output in order to represent the probabilistic score of each category. This system has an added advantage of resilience against missing data thanks to its dropout layer, which is trained to work when data from a source may no be available or reliable. For example, in cases where the image model cannot decide on any category, the dropout layer would give more weight to the input from the text based models.

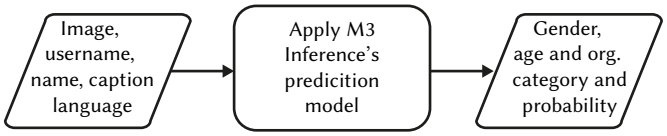

Fig. 1. M3 Inference's input and output.

B. Flesch Reading Ease

The *Flesch Reading Ease* is a simple and effective metric to measure readability invented in 1948 by Rudolph Flesch, which can be adapted to several languages [32]. Such expression (equation 1) gives a score related to the difficulty of the text analyzed, drawing a value within the range of 0 to 100 as a result. The lower readability score, the higher the level of education that the transmitter is supposed to have.

(1)

Table I shows the distribution of education levels according to this indicator. Therefore, a text made with short sentences and simple words would have a very high score, and a text made up of long and complex sentences would have a small value assigned.

TABLE I. Flesh Reading Ease Score Distribution

Flesch Reading Ease Categories	
Score	Level of difficulty
90-100	5th grade - Very easy
80-90	6th grade - Easy
70-80	7th grade - Fairly easy
60-70	8th to 9th grade - Plain English
50-60	10h to 12th grade - Fairly difficult
30-50	College - Difficult
10-30	College graduate - Very difficult
0-10	Professional - Extremely difficult

C. Word2Vec Model

In the area of natural language processing, words are generally interpreted as associated symbols that do not necessarily have to have meaning; thus, the word *tourism* could have an associated id such as *id1*. The problem of following this nomenclature lies precisely in not keeping a relationship between similar terms, and it is for this reason that space vector models have arisen, pretending to group words that belong to the same lexical family.

From this idea was born the Word2Vec model, which is mainly found in two forms: Continuous Bag-of-Words (CBOW) or Skip-Gram. The former consists of trying to determine the word in the middle of a text from the context taken from the rest surrounding the word of interest, and is the one used for the present work. On the other hand, the latter is based on trying to predict context of the text based on a given word of interest.

The way to do this is by using a technique called one-hot vector, which is based on creating a vector with as many zeros as words in the text and assigning a 1 to the position where the word of interest is located. Thus, for instance, if you have the sentence *Today is sunny* and you want to encode the word sun, the vector would look like [001].

The output of the Word2Vec Model is a vocabulary in which each item has an associated vector that can be used to identify similarities and relationships between all the words that characterize the images

from instagram, which in this case is composed of a 300-dimensional Word2vex model.

D. Principal Component Analysis

After completing the word encoding, it was decided to perform a step to reduce the dimensionality of the data from 300 to 2-dimensional, in order to both facilitate and optimize its subsequent classification. It is important to note that technically one dimension would have been sufficient to explain the variability of the vectors (since the first component is able to explain 90% of the variability); however, two have been chosen in order to follow the recommended methodology and provide greater representation and explanation of the differences in the data. *Principal Component Analysis*, i.e. PCA, is a reduction dimensionality technique used to improve the understanding of a large dataset, minimizing the information loss by creating new variables that are not correlated [37] [38]. In this way, enough components have been retained to create a two-dimensional space, which explains or contains more than 90% of the variance of the data.

E. Unsupervised Clustering Algorithm: KMeans

Once the reduction of the vector space corresponding to the tags that were taken from the instagram images is completed, the next step consists of applying one unsupervised algorithms known as KMeans. The term unsupervised implies that the input data does not contain labels, so the clustering is performed by similarity in the vectors representing the chosen words.

A parameter k is defined, which refers to the number of centroids to be searched for in the dataset, i.e. the number of clusters to be obtained. These centroids correspond to the center of the cluster in question. Once the number of clusters is selected, the algorithm starts iterating to optimize the position of the centroids of each cluster. In such a way, it will only stop if the centroid position stabilizes after a number of iterations, or if a maximum number of iterations is reached.

1. Silhouette Score

Aiming to identify the optimal number of clusters to be chosen for the KMeans clustering algorithm, the average silhoutte method is evaluated in order to determine how well each item is classified within its cluster. The higher the average silhouette width, the more appropriate the item is classified. Calculations for different numbers of groupings show that the optimum is two, since it corresponds to a value of 0.76 according to the silhouette method, being the highest of all the tests performed (Table II).

TABLE II. Silhouette Method Coefficients

N° clusters	2018	2021
2	0.7576	0.7699
3	0.6696	0.6781
4	0.6392	0.6484
5	0.6124	0.6268
6	0.5729	0.5640
7	0.5707	0.5658

III. Results

This section describes the data used in this study, discusses the challenges in collecting it and explains its format as a preamble. The enrichment process and its components are described, what builds up the inference pipeline, and the different machine learning approaches used are discussed - directly or indirectly - as well as their strengths. An extra section will describe the activity classification model that has been developed for this study, explaining the different stages and techniques required to build it.

A. Data Description and Enrichment

The data used in this study is a set of 10,000 Instagram posts tagged in Vigo, Spain [39] and comprises two different datasets: 5,000 posts from the last two weeks of May 2018 and 5,000 posts from the last two weeks of May 2021. The goal of dealing with these datasets is to try and find differences in visitors - namely, people posting at a given location - before an after an event expected to have impacted travellers' and visitors' behaviour, such as the SARS-CoV-19 pandemic. It is important to note that, while desirable, obtaining large amounts of posts from Instagram is not an easy feat, thus efforts to provide inference tools from low amounts of data are required.

Table III shows an example of the data used to conduct the study in table format. More information could be retrieved, but these were the only parameters our enrichments needed.

TABLE III. Data Format Example

image_url	https://instagram...
caption	Enjoy these moments...
username	the_foobaz
full_name	Eugenio Doe

Our data processing pipeline is shown in Fig. 2 as a set of independent processes that infer the different dimensions of interest. This process is referred as the enrichment of the data set because it infers valuable information from *raw* data. The overall pipeline consists of a wide range of machine learning and data mining techniques combined with the usage of external MLaaS1 platforms that allow us to get some promising results with small sets of data.

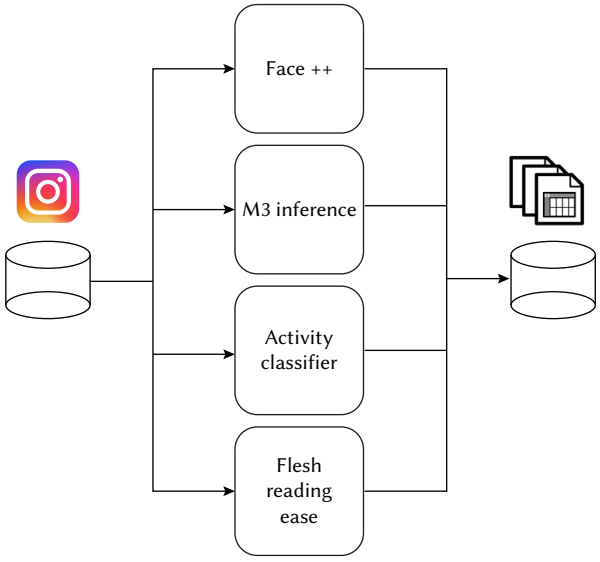

Fig. 2. Enrichment on data

To obtain demographic information two machine learning methods are utilized: Face++ platform [30] and M3Inference [29]. The first uses facial recognition to identify faces and estimates their demographics based on them, while the second analyses the image and text data from the user to make the same inference. As explained before, the latter is used to obtain information where the former fails to recognize any face. The activity dimension is also inferred using machine learning and data mining techniques described later.

A different approach, however, is taken for the educational dimension, where instead of relying machine learning models the Flesch Readability formula was used. One such approach became the seed for more sophisticated evaluations, and even adaptations

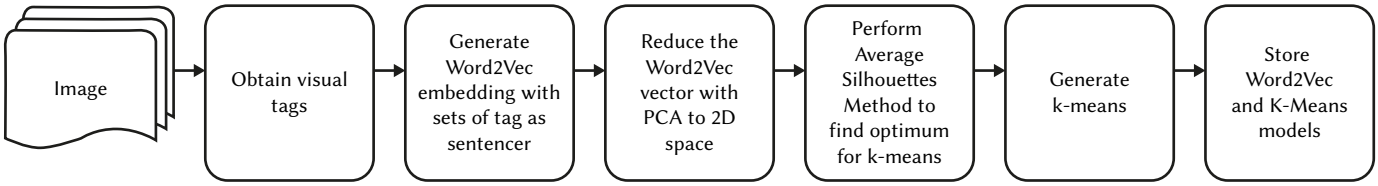

Fig. 3. Activity classifier generation process.

in different languages. Moreover, it offered flexibility on multiple languages -making switching between them straight forward thanks to the python implementation found in [40].

Some of these systems are closed source, such as Face++ and Azure's Cognitive Services [41] so no details on their techniques can be discussed, but provided their services cover the areas of facial and object recognition, it can assumed there's a high degree of deep or convolutional neural networks at work given their outstanding performance on image classification tasks [42]. That is the case for one of the ways the demographic dimension of a subset of posts is obtained from the data set through Face++'s face detection web API, which searches for faces in images, analyses them and infers their age and gender, among other things, and returns that to the caller.

B. Clustering

1. Activity Classifier. Model Generation

Based on the work of [24] we made our own implementation of the activity classifier with the data we had. The overall process of generating the classifier is shown in Fig. 3, which mentions the techniques used in this approach.

The main parts of the process are explained as such:

1. Obtain visual tags from each image with Azure's computer vision API [41] and make sentences by joining all of the tags into text. These tags can be anything that the API recognizes, such as: outdoors, beach, water, sand, bikini.
2. Train a *Word2Vec* [43], [44] with the sentences to generate a word embedding that learns relationships between words. For example, words that repeatedly appear in the same sentences together will be more similar than those that do not.
3. Reduce the vector space from the *Word2Vec* model using PCA [38] to project the two main components - which provided the highest variance - into a new 2D space in order to allow for more efficient clustering. Other approaches were tested, such as UMAP [45] and T-SNE [46], but after comparing the final clusters by means of image sampling and by the sets of words that defined them, it was concluded that PCA offered similar results in significantly less time. On T-SNE, it was decided not to be a good option to make a reusable model because it is not deterministic, therefore there is no guarantee that new similar data would be classified into the right clusters.
4. Compute the Silhouette Score [47] for different values of k to discover the most efficient value for the K-Means clustering algorithm.

2. Data Classification Interpretation

Fig. 4 shows the comparison between the age profiles found within each grouping. First of all, it is noteworthy that the first cluster contains a greater number of individuals than the second cluster, which is composed mostly of a female profile. The age focus is centered between 25-30 years for both clusters, showing a downward trend towards older ages and no difference pre- and post-pandemic. On the other hand, for the second of the clusters a greater comparison between the profiles can be seen, counting a lower presence of women for 2021. The center of age shifts more towards 30 years of age and the downward trend seen in the first cluster is lost.

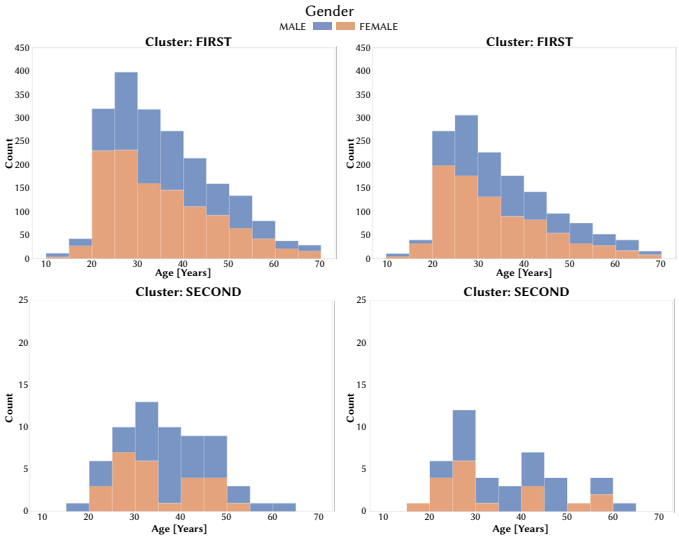

Fig. 4.Age comparison between clusters from 2018 (left) and 2021 (right)..

Moreover, Fig. 5 shows a box plot for each year, comparing the difference in the education profile of the publications between the first and second clusters, taking into account gender. In the first of the clusters, a higher level of Flesch's pre-pandemic indicator in men can be seen, which is equivalent to a lower difficulty of the analyzed text. This tendency is also seen to a lesser extent in the case of women, who present a greater dispersion of the data and an average that remains around 50 in 2018, decreasing towards 40 by 2021. In the case of the second cluster, it is clear that for both women and men, the level of difficulty of the texts analyzed increased post-pandemic.

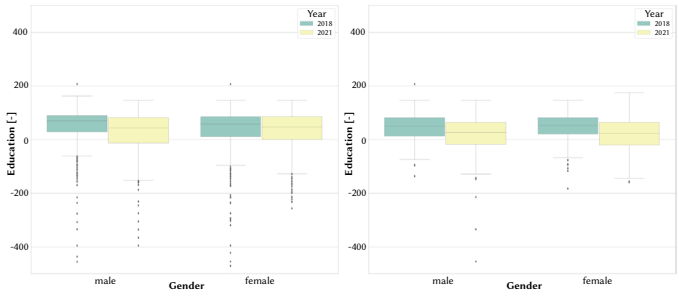

Fig. 5. Age comparison between years from first (left) and second (right) clusters.

IV. Conclusions and Future Lines

This work has developed a methodology for analysing tourism through the prism of complex mathematical algorithms, based on unstructured data extracted from social networks.

As a conclusion of this work, it is possible to appreciate the great potential offered by the analysis of information from social networks for the identification of tourism profiles. Serving in this case to locate differences between people visiting the city of Vigo between the years 2018 and 2021. In the results, a more notable difference has been observed in the education levels of the profiles, rather than in the age ranges; being logical that most of them focus on lower ages where social networks are more successful.

Beyond that, the geolocated image extraction and analysis methodology implemented in the present work has great potential for comparisons on larger amounts of data and even between tourism profiles between cities.

Once the results obtained in this work have been analysed, new algorithms are being developed that integrate several data sources, as well as the improvement of enrichment techniques, always oriented towards decision-making in the tourism sector.

Acknowledgment

Pilar Muñoz has received support from the Spanish ministry of Science and Research (grant PID2020-116040RB-I00). The work of Ana Larrañaga has been supported by the 2020 predoctoral grant of the University of Vigo.

References

[1] J. G. Brida, S. London, M. Rojas, "El turismo como fuente de crecimiento económico: impacto de las preferencias intertemporales de los agentes," *Investigación económica*, vol. 73, pp. 59 – 77, 09 2014.

[2] I. Cortés-Jiménez, "Which type of tourism matters to the regional economic growth? the cases of spain and italy," *International Journal of Tourism Research*, vol. 10, no. 2, pp. 127–139, 2008, doi: https://doi.org/10.1002/jtr.646.

[3] W. T. Organization, "International tourism highlights," 2019.

[4] U. N. W. T. Organization, "Unwto global tourism dashboard. country profile - outbound," 2020.

[5] W. Travel, T. Council., "Research – economic impact reports.," 2020.

[6] A. Santana Talavera, "Patrimonios culturales y turistas: Unos leen lo que otros miran," *PASOS : Revista de Turismo y Patrimonio Cultural*, vol. 1, 01 2003, doi: 10.25145/j.pasos.2003.01.001.

[7] F. Jiménez, C. y Seño, "Patrimonio cultural inmaterial de la humanidad y turismo.," *International Journal of Scientific Management and Tourism*, vol. 4, no. 2, pp. 349– 366, 2018.

[8] G. Yudice, "El recurso de la cultura.," *Gedisa. Barcelona*, 2001.

[9] UNESCO, "Convención para la salvaguarda del patrimonio cultural inmaterial de la unesco.," 2003.

[10] J. Arévalo, "La tradición, el patrimonio y la identidad," pp. 925–955, 2004.

[11] M. Timón Tiemblo, M.P. y Domingo Fominaya, "Resumen del plan nacional de salvaguarda del patrimonio cultural inmaterial," *Anales del Museo Nacional de Antropología*, vol. 14, pp. 29–44.

[12] J. Nared, D. Bole, *Participatory Research on Heritage- and Culture-Based Development: A Perspective from South-East Europe*, pp. 107–119. Cham: Springer International Publishing, 2020.

[13] B. A. Adie, C. M. Hall, "Who visits world heritage? a comparative analysis of three cultural sites," *Journal of Heritage Tourism*, vol. 12, no. 1, pp. 67–80, 2017, doi: 10.1080/1743873X.2016.1151429.

[14] C. Milano, M. Novelli, J. M. Cheer, "Overtourism and tourismphobia: A journey through four decades of tourism development, planning and local concerns," *Tourism Planning & Development*, vol. 16, no. 4, pp. 353–357, 2019, doi: 10.1080/21568316.2019.1599604.

[15] P. L. Winter, S. Selin, L. Cerveny, K. Bricker, "Outdoor recreation, nature-based tourism, and sustainability," *Sustainability*, vol. 12, no. 1, 2020, doi: 10.3390/su12010081.

[16] Y. Deng, C. Li, "Research progress, theories review and trend forecast on placeness of tourism destination," in *Proceedings of the 3rd International Seminar on Education Innovation and Economic Management (SEIEM 2018)*, 2019/01, pp. 431–434, Atlantis Press.

[17] R. E, "Classics in human geography revisited, place and placelessness.," *Progress in Human Geography*, vol. 24, no. 4, p. 613, 2000.

[18] S. A. Bowen, D, R. E, "Tourist satisfaction and beyond: tourist migrants in mallorca.," *International journal of tourism research*, vol. 10, no. 2, pp. 141–153, 2008.

[19] A. Bowen, "War-affected children in three african short stories: Finding sanctuary within the space of placelessness.," *Commonwealth Essays and Studies*, vol. 42, no. 2, 2020.

[20] T. Wenyue, "The influence and significance of tourism development on placeness.," *Tourism Tribune*, vol. 28, no. 4, pp. 9–11, 2013.

[21] L. Leilei., "The spatial cognition process and law of tourist destination image.," *Scientia Geographica Sinica*, vol. 6, pp. 563–568, 2000.

[22] W. B, "Regional tourism plannig principles.," *China Travel and Tourism Press*, 2001.

[23] K. X. Zhou S Y, Yang H Y, "The structuralistic and humanistic mechanism of placeness: A case study of 798 and m50 art districts.," *Geographical Research*, vol. 30, no. 9, pp. 1566–1576, 2011.

[24] G. Kalra, M. Yu, D. Lee, M. Cha, D. Kim, "Ballparking the urban placeness: A case study of analyzing starbucks posts on instagram," in *International Conference on Social Informatics*, 2018, pp. 291–307, Springer.

[25] M. K. . M. F. Pfeffer, J., "War-affected children in three african short stories: Finding sanctuary within the space of placelessness.," *Commonwealth Essays and Studies*, vol. 7, no. 1, p. 50, 2018.

[26] B. E. Rossi, L., A. Torsello, "Venice through the lens of instagram: A visual narrative of tourism in venice.," *Companion Proceedings of the The Web Conference 2018*, pp. 1190–1197, 2018.

[27] K. Jang, Y. Kim, "Crowd-sourced cognitive mapping: A new way of displaying people's cognitive perception of urban space.," *PLoS ONE*, vol. 14, no. 6, p. e0218590, 2019.

[28] U. S. Hasan S, Zhan X, "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media.," *PLoS ONE*, vol. 14, no. 6, p. e0218590, 2003.

[29] Z. Wang, S. Hale, D. I. Adelani, P. Grabowicz, T. Hartman, F. Flöck, D. Jurgens, "Demographic inference and representative population estimates from multilingual social media data," in *The world wide web conference*, 2019, pp. 2056–2067.

[30] Beijing Kuangshi Technology Co., Ltd., "Face++ platform." https://www.faceplusplus.com/ face-detection/, 2021. Accessed: 2021-07-22.

[31] G. d. Q. J. B. S. Alvarez, P., "Riada: A machine-learning based infrastructure for recognising the emotions of spotify songs,.," *International Journal of Interactive Multimedia and Artificial Intelligence. IN PRESS*, 2022.

[32] R. Flesch, "A new readability yardstick.," *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.

[33] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[34] W. C. W. T. I. W. K. P. C. Y. H. . H. K. S. Chen, S. H., "Modified yolov4-densenet algorithm for detection of ventricular septal defects in ultrasound images.," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 7, pp. 101–108, 2022.

[35] S. Hochreiter, J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, ch. 6, pp. 180–184. MIT Press, 2016. http://www. deeplearningbook.org.

[37] I. T. Jolliffe, J. Cadima, "Principal component analysis: a review and recent developments," vol. 374, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[38] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[39] Instagram, "Vigo, spain on instagram • photos and videos." https://www.instagram.com/explore/ locations/23436873/vigo-spain/. Accessed: 2021-07-25.

[40] S. Bansal, C. Aggarwal, "textstat | pypi." https:// pypi.org/project/textstat/, 2021. Accessed: 2021-07-29.

[41] Microsoft Corporation, "Computer vision | microsoft azure." https://azure.microsoft.com/es-es/services/cognitive-services/computer-vision/#overview, 2021. Accessed: 2021-07-26.

[42] C. Szegedy, A. Toshev, D. Erhan, "Deep neural networks for object detection," 2013.

[43] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space," 2013.
[44] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality," 2013.
[45] L. McInnes, J. Healy, J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020.
[46] L. Van der Maaten, G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.
[47] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

Pilar Muñoz Dueñas

Pilar Muñoz Dueñas is a tenured Professor at the University of Vigo. She holds a doctorate in Financial Economics and Accounting from the University of Vigo. She has taught both undergraduate and postgraduate courses. She has written several publications and articles. She is the main researcher of the Interreg Atlantic CultureSpace project and she participates as a researcher in others at international (NTERREG POCTEPT) and national level (Retos of the Spanish Ministry of Science and Innovation).

Eugenio Doñaque Gonzalez

Eugenio Doñaque Gonzalez is a software developer and a student of informatics engineering at the Open University of Catalonia. He is currently working in the financial technology sector, and has experience with data processing systems and applied data mining and machine learning.

Ana Larrañaga Janeiro

Ana Larrañaga Janeiro is a graduate in Energy Engineering from the Universidade de Vigo (2019). She received an Interuniversity MsC in Industrial Mathematics (M2i) from the Universidade de Vigo, Santiago, Coruña, Politécnica de Madrid and Carlos III (2021), and is currently a predoctoral researcher at the Center for Research in Technologies, Energy and Industrial Processes of the same university (CINTECX), where she focuses her scientific research in the areas of Artificial Intelligence and Computational Fluid Dynamics (CFD). She focuses her research on the application of Machine Learning techniques to improve Computational Fluid Dynamics (CFD) calculations, the subject of her doctoral thesis in the Interuniversity Doctoral Program in Energy Efficiency and Sustainability in Engineering and Architecture at the University of Vigo.

Javier Martínez Torres

Javier Martínez Torres is a Mathematician and Engineering PhD from the University of Vigo. He is currently an Assistant Professor at the University of Vigo and has participated in more than 20 research projects as principal investigator. He has published more than 60 papers in JCR indexed journals and participate in more than 35 international conferences.

Ana M. Mejías

Ana M. Mejías received her PhD in Industrial Engineering from The Universidad Politécnica de Madrid (UPM-Spain). She is an Associate Professor at the University of Vigo (Spain) and she is Vice Dean in the School of Industrial Engineering. She has participated in more than 10 research projects and she has a patent in exploitation; She has published 30 papers in indexed journals and participate in more than 50 international conferences.