Online Social Networks and Media 37-38 (2023) 100274

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.journals.elsevier.com/online-social-networks-and-media

Reputation assessment and visitor arrival forecasts for data driven tourism attractions assessment

Enrico Collini, Paolo Nesi *, Gianni Pantaleo

Distributed Systems and Internet Technologies Lab, Department of Information Engineering, University of Florence, Florence, Italy

ARTICLE INFO

Keywords:
People count prediction
Reputation computing and predictions
Data driven tourism attraction management

ABSTRACT

Tourism is vital for most historical and cultural cities. In the context of Smart Cities, there are numerous data sources in tourism domain that could be analyzed to monitor and forecast a range of different indicators related to touristic locations and attractions. In this paper, we propose a framework which exploits social media and big data to forecast both online reputation and touristic attraction presences. To this end, some techniques have been tested and proposed on the basis of machine learning, deep learning, causality assessment and explainable Artificial Intelligence, so as to provide evidence of the relevant variables for each prediction and estimation. An approach has been introduced to analyze the explainability of the proposed solutions, i.e., a multilingual sentiment analysis tool for social media data based on transformers to compare data sources as Trip Advisor and Twitter. Furthermore, causality analysis has been performed to evaluate the temporal impact of social media posts and other factors with respect to the number of presences. The work has been developed in the context of Herit-Data, a European Commission funded project on the exploitation of big data for tourism management and based on the Snap4City infrastructure and platform. Herit-Data has developed solutions for 6 major European touristic locations. In this paper, some of the solutions developed for Florence, Italy and Pont du Gard, France, are reported.

Introduction

Tourism is perceived as an experience which needs to be communicated [1] and social media is the perfect instrument and communication channel to implement the so-called *word of mouth* [2]. Social media content is gaining more and more relevance in the field of tourism, because communities and mechanisms for the propagation of information are perfect vehicles for web reputation on attractions and cultural heritage sites.

Marchiori et al. in [3], has outlined that one of the main directions in touristic domain was understanding how the perception analysis of online destination content could influence tourists' destination choice. The success of touristic destinations is influenced by their reputation [4]. Arumugam et al. in [5] has framed this concept as Destination Branding whose improvement could lead to more visitors, thus more financial gains. More tourists can further promote engagement and increase the online reputation of the location. Platforms such as Trip Advisor collect comments from their users and the provided scores in terms of stars can be regarded as explicit reputation assessment and promotion. The reached scores are presented to the audience in the form of average scores, to motivate attendance and to push the user to read comments; thus, tourists in most cases decide to visit or avoid that destination on such a basis. In [6], authors have argued that tourists are keener to select destinations upon their reputation, and for this reason reputation assessment is considered to be a major asset in tourism business. Metrics for such assessment of tourism destinations have been proposed in [7]. Notably two different categories of reputation measurements have arisen: general feedback ratings, such as the score metric of TripAdvisor, and derived/composed indexes, as reported in [8] about other online data resources for reputation assessment issue.

In [9], an Online Reputation Management has been proposed exploiting social media data analysis where tourists share their experiences (both positive and negative) about attractions. Therefore, works in social media sentiment analysis are becoming increasingly relevant [10], thus representing an indirect measure of reputation. Moreover, there is a large variety among social media platforms in terms of reaction time by their users. Platforms such as Trip Advisor engage users via reviews and scores with a certain delay, with respect to the visit time;

* Corresponding author.
E-mail address: paolo.nesi@unifi.it (P. Nesi).

https://doi.org/10.1016/j.osnem.2023.100274
Received 4 August 2023; Received in revised form 14 October 2023; Accepted 28 October 2023
Available online 8 November 2023

others such as Twitter allow a faster spreading of information, letting people communicate and share content more quickly. Twitter is one of the most responsive social media platforms and may be used to convert community sentiment into numerical metrics by using specific tools, such as [11]. Assessing the sentiment of both tweets and comments, in general, can be of extreme importance to understand if they are leading users towards a positive discussion or towards dissatisfaction. The predictive capabilities of Twitter in combination with sentiment analysis algorithms allow the development of solutions to assess a tweet's orientation, appreciation of services [12], recommendations for places to visit [13], and it can be used in combination with other additional pieces of information from big data sources to forecast the online reputation of a tourism destination. The volume of posts/comments generated by rapid social media activity in platforms such as Twitter has been also used to estimate and forecast the attendance at events and other kinds of people presences even in other scenarios outside the tourism field, such as TV shows [11], and tourist demand [14], mainly with regression models such as seasonal autoregressive integrated moving average SARIMA [15]. Users' tweets are displayed to each and everyone of his/her followers and they can retweet the original tweet echoing the message to their own followers, as well. These high volumes related to specific keywords and hashtags can represent, when monitored, useful sources of information for the development of multiple applications for decision makers, computing forecasts of event aspects such number of people and they can be also used for early warning systems and service tuning.

In fact, tourist location capacity is a multidimensional value depending on contextual data and information like environmental conditions that may influence people presences at tourism destinations. The capacity of a specific location or service can vary in different periods of the year. Other aspects which may affect the number of presences in touristic attractions could be related to weather conditions. This means that a framework, in order to produce predictions and data analytics in the touristic domain, should collect heterogeneous data coming from social media, weather status and forecasts, local sensors such as counting of tickets and presences, etc. In particular, social media analysis implies to perform both natural language processing (NLP) and sentiment analysis (SA), while some time-series analysis may be needed for other big data sources.

A deeper description of the paper's aims, and structure is reported in Section II, while the following Section I.A illustrates the state-of-the-art analysis and related work. As a preview, in this paper, we propose a framework mainly focused on reputation and related effects to help decision makers to estimate and predict reputation on the basis of data which can be recovered in real time from social media and other sources. This framework highlighted the impact of some specific variables in predicting reputation and presences. In addition, a causality analysis has been provided to assess the impact of reputation, presences and posts on social media in the predictions of presences. The motivations lying behind any reputation and/or presence prediction are connected with the expenditure prediction in that area and this may improve any tourist experience and market strategies (which also implies to understand the cause effect phenomena from promotion on social media to the number of presences). The proposed solution exploits social media and other big data.

State of the art and related works

There is more and more emphasis on providing decision makers with predictive tools related to tourism services. These tools focus on two main indicators: online reputation and number of presences in touristic attractions. By developing predictive models based on data-driven approaches, such as machine learning and data analytics, researchers seek to empower decision makers with the ability to both anticipate changes in reputation and forecast visitor presences.

Reputation assessment

Numerous research papers have addressed the topic of Tourism Reputation Assessment, with a particular focus on online platforms. Some of these studies have investigated how social media content on online tourism platform influences the perception of attractions, aiming to identify the most influential platforms for tourism reputation management. For instance, in [16], authors compared three major platforms in the hospitality and tourism industry in Manhattan, New York City. These platforms incorporated tourists' reviews as primary social knowledge, and the study has reported that, despite their differences in various aspects, Trip Advisor stood out with higher overall quality for reputation assessment. This finding helps explaining why Trip Advisor has been widely perceived as a premier data source. Other studies focused on the relationship between review sentiments and content on a destination reputation. For example, in [17], authors have performed text mining and sentiment analysis of Trip Advisor reviews to understand their emotional polarity as conveyed in the addressed reviews and main topics. On the other hand, Trip Advisor is unsuitable to build an Online Reputation Management system, because of its delay from visit time and collection of reviews. Predictive capabilities of Twitter have been exploited in [11–13], thus enabling the acquisition of tourism-related tweets via specific tourism-related keywords. In the context of reputation assessment, the relationships between Twitter, TW, and TripAdvisor have not been analyzed in literature, so the results reported in this paper could provide some evidence on these aspects.

Reputation forecasting

Decision makers could benefit from Online Reputation Management systems to assess the online reputation of touristic attractions and even more to identify which are the drivers increasing the reputation of attractions, and thus any possible reputation prediction. This topic is quite complex, similar to the prediction of stock options values. Puh et al. [18], have proposed a solution for reputation predicting by using Trip Advisor review data. The idea behind both reputation prediction and volume of presence prediction is related to the draft prediction of expenditure, improving the tourist experience and market strategies.

Moreover, reputation changes over time and can be regarded as a time-series data. In this context, the field of time-series prediction has attracted considerable scientific attention, especially with a focus on providing frameworks and artificial intelligence (AI) models and architectures, so as to achieve state-of-the-art results for specific applications. The state-of-the-art solutions on AI use methodologies such as ensemble-learning techniques, including random forest (RF) and gradient boosting machines (XGBoost) [19], and Deep Learning specifically for time-series forecast. In this field recently two types of architectures have been reported to achieve better performances with respect to previously used recurrent neural networks (RNNs): the temporal convolutional networks (TCN) [20], and the temporal fusion transformers (TFT) [21].

Tourist presence forecasting

The range of applications for predicting the number of people in various contexts has fueled extensive research. Within the touristic domain, predictive capability carries value for decision makers and there is a range of works dealing with people presence prediction in touristic attractions, on the basis of several kinds of data; a selection of such works is reported in Table 1. Data sources could range from devices for counting people (paxcounters) which can be based on Wi-Fi sniffing [22], laser counting [23], to thermal cameras [24], Twitter data [11,12].

Ivanovski et al. in [25] have applied a moving average model for the quarterly prediction of the number of tourists in the Republic of Macedonia. The provided data ranged since 2012 till 2017 (historical data on presences), and authors have used a univariate approach to forecast the number of tourists achieving a mean absolute percentage error (MAPE) of 6.45 % on the test set of the last 2 quarters of 2017.

Chang et al. in [26] have compared SVM (support vector machine)

Table 1
Related work on Tourism Presences Forecasting.

Authors	Target	Features	Dataset	Model	Results			
Ivanovski et al. 2018 [22]	Prediction on the number of tourists	Number of tourists per quarter	Republic of Macedonia	Moving Average (MA) Model	Method	MAPE test set Q1	MAPE test set Q2	
					MA	9,49%	5.61%	
Chang et al. 2017 [23]	Visitor Arrivals	68 Features from Governmen Website – then Feature Selection	Taiwan	SVM, DNN	Method	MAPE		
					SVM	3.12%		
					DNN	2.05%		
Laaroussi et al. 2020 [24]	Tourism Demand forecasting	Month -Year Domestic and Foreigner Total of tourists	The Tourism Observatory tourist arrival	LSTM, GRU, ANN, SVR (SVM) 1 month forecast	Method	MAE	MSE	MAPE
					LSTM	35.19	857.49	9.23%
					GRU	36.77	964.41	9.81%
					ANN	43.15	1576.98	12.65%
					SVR	39.70	1420.81	11.45%
Chen et al. 2022 [25]	Tourism Demand Forecasting	Tourist Volume, Air, Climate, Holidays, Baidu Index, and News Coverage	Jiuzhaigou	KNN, XGBOOST, LSTM	Model	MAE	RMSE	
					KNN	541.030	971.106	
					XGBOOST	533.173	936.087	
					LSTM	547.210	983.498	
Li et al. 2020 [29]	Tourism Demand Forecasting	Weekly tourist arrivals, search queries, Online reviews	Mount Siguniang, China TOT 132 WEEKS of data	ARIMAX, SVM, RF 1 week forecast	Model	MAE (week)	MAPE	
					ARIMAX	4417.90	43.47%	
					SVM	3668.12	26.72%	
					RF	3491.26	4.09%	

and Deep Neural Network (DNN) to forecast tourist arrivals in Taiwan with a dataset of 69 features primarily gathered form the Government Website. Authors have exploited a multivariate approach to predict tourist arrivals, and a feature selection has been applied, too, in order to improve overall results on the test set. Results reported a MAPE of less than 10 %. Authors stated that forecasting visitor arrivals is of great importance, since it is an indicator of tourism demand and can serve as a reference for governmental policies about tourism and business strategies to be conceived by tourism industries.

Laaroussi et al. [27], have assessed the issue of tourism demand forecasting, specifically for the Moroccan tourism observatory on monthly arrivals. Authors approached the work using multivariate data sources, including datetime and number of tourists (domestic, foreign, and total) as features. Laaroussi et al. have compared different types of architectures: ANN (Artificial Neural Network [30]), SVR (Support Vector Regressor [31]), GRU (Gated Recurrent Unit [32]), and LSTM (Long Short-Term Memory networks [33]) which is a recurrent Deep Learning architecture which turned out to be the best performing one. As reported before, currently there are other architectures that are achieving better results than the above-mentioned models. Phan et al. [34], have compared the TCN with the recurrent neural networks such as LSTM and GRU, and also with ensemble learning techniques such as XGBoost (which has proved to be a valid solution to be taken into consideration in this kind of tasks). Another type of architecture that is gaining popularity for its results is the TFT, specifically for time-series forecasting and modeling problems. Hu et al. in their work [35], have tested this type of forecasting architecture, by comparing it with SVR and LSTM. Their work has mainly focused on time-series forecasting and the transformer-based architecture has achieved better results with respect to SVR and LSTM.

Chen et al. have focused their work [28] on the use of heterogeneous data sources for tourism demand forecasting applications. Their case study used data with daily time granularity and considered tourist volume, air, climate, holidays, Baidu index, and news coverage for tourism demand forecasting in Jiuzhaigou, with a data history since December 25, 2015, up to January 7, 2022. Researchers highlighted the importance of including data sources related to tourism sentiment classification. This is especially important in tourism demand forecasting applications, because "sentiment classification revealed the emotional tendencies of the social events hidden behind the text".

Puh et al. have focused their work [18] on extracting sentiment and ratings from tourist reviews and have agreed that these data sources can provide a lot of useful information to be used to greatly improve touristic experiences.

Li et al. [29], faced the problem of tourism demand forecasting for Mount Siguniang, in China, by using multiple data sources including data related to online reviews, thus remarking the importance of these types of information, same as in the works of Chugh et al. [36], and Puth et al. [18]. Researchers could compare different solutions for the problem of weekly tourism demand forecasting on Mount Siguniang in China: ARIMAX [37], SVM [31], and random forest, RF [38] for multiple time horizons (1,2,3,6,9,12 weeks). Results have demonstrated that the ensemble learning technique performed better for the time horizon of 1 and 2 weeks; besides, SVM achieved better results for the longer time horizons of 3, 6, 9, and 12 weeks. The main finding is related to the improvement of results when including data concerning online reviews.

Song et al. [39], have reported useful insights on how to predict tourism demand using Big Data. Researchers demonstrated that social media data from micro-blog platforms like Twitter can provide useful insights into tourism opinions and behaviors, and therefore they should be included in any analysis on tourism metrics forecasting. The proposed framework stresses the importance of a human-centered artificial intelligence (HCAI) approach where, as researchers stated, "*…we need to have a deep understanding of where the data come from, what audience will be consuming the data, and how that audience will interpret the information.*". Miah et al. [40], have reported that Big Data from social media sites is a relevant source of data to gain a better understanding on user behaviors, which can be of value for decision makers. In this case, social media as Flickr have been used to extract and process social media posts with text, images and geolocations. Outcomes have been the identification of interests and the volume of posts associated with specific sites over time. On this latter aspect, predictions have been produced. Therefore, structured and unstructured data have generated useful information on the tourist's behavior and could support the results obtained on the case study in Melbourne, Australia.

Article contribution and structure

This article contributes to the work of decision makers in the domain of tourism management by proposing an integrated framework for reputation prediction by using data analysis and data-driven approaches. Assessing online reputation is not enough to enable decision makers to monitor the impact of actions on tourism attractions. Actually, the DSSs need to be capable to forecast online reputation metric and/or the number of presences, as well as understanding drivers. The results generated by data driven processes could be exploited by decision makers to improve the quality of the service towards any assessment of a tourism destination situation, and they could be a good instrument to understand the effects of events, advertising, etc., once accompanied by specific explanations and analyses. Therefore, the main objectives of this research are related to presenting an integrated framework mainly focused on reputation and related effects, which allows to perform the:

- **Estimation and prediction of reputation** as to touristic attractions. To this end, an analysis on social media posts and volume of presences has been considered, by comparing Twitter (TW) and trip advisor (TA) Data. The adopted technique has been NLP (Natural Language processing), SA (Sentiment Analysis), AI and eXplainable AI (XAI). The XAI techniques allowed to identify the main features to be tracked to compute the reputation predictions.
- **Long term prediction of presences** attending touristic attractions. In this case, a larger set of data sources has been used, while taking into account social media, weather and historical data. The prediction of presences has to be long term to let decision makers consider any possible insight, for example, service and security tuning of the attraction. In this case, the prediction of presence has been analyzed in terms of causality.
- **Causality analysis** allowed to understand drivers and motivations to attract tourists in specific locations. Therefore, the analysis has been performed on the same data using the predicting presences and it led to discover that the most relevant features impacting on presences are both reputation and the effects of social media posts with a delay of more than 15 days.

To sum up, the research aimed to identify a framework based on Big Data and social media providing tools for reputation assessment and prediction, prediction of presences and causality analysis, taking into account social media data, Twitter and a set of big data, that may be accessible for cultural heritage sites.

The work presented in this paper has been developed in the context of Herit-Data Interreg project of the European Commission, with the target of exploiting Big Data and artificial intelligence for tourism management. Herit-Data has involved 6 major touristic areas: Florence in Italy, Pont du Gard in France, Valencia in Spain, Mostar in Bosnia and Herzegovina, Dubrovnik in Croatia and Ancient Olimpia in Greece [https://herit-data.interreg-med.eu/]. Relevant attractions share the same needs of monitoring presences, predicting presences and reputation over time, analyzing social media data for such a purpose. To this aim, Tweets in multiple languages (i.e., English, Italian, Greek, French, Spanish) have been collected and processed, by using Twitter Vigilance platform [11], as a basis for computing a number of Tweets based on volume and Sentiment Analysis metrics. In addition, data coming from sensors and TripAdvisor have been collected as well.

The general data/process flow concerning the exploitation of the mentioned data to perform the analysis and for computing predictions and reputations is reported in Fig. 1, also referring to the paper sections and subsections.

As major validation experiments, the choice has been to work on data from Florence and Pont du Gard locations and their related major attractions. Both reputation and touristic presences at the Uffizi Gallery in

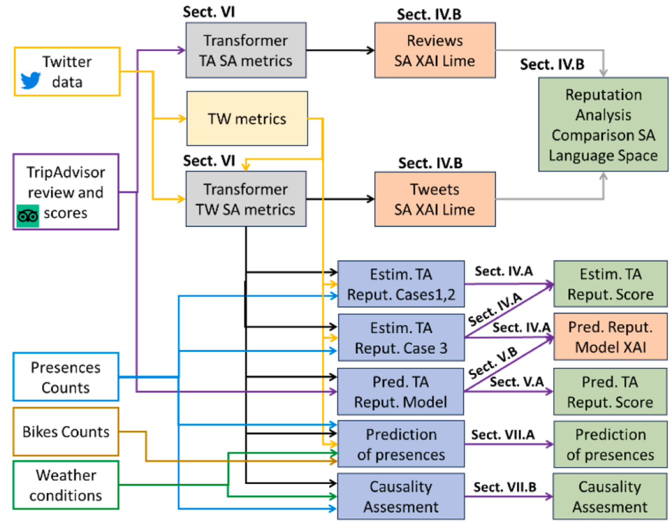

Fig. 1. Paper structure with data flows and processes (colored blocks). The section numbers indicate where the related issues are discussed in the paper. The whole suite of data has been used in almost all the experiments, and in the figure only the most relevant connections are reported.

Florence (Italy), which is one of the most visited museums in Europe, have been explored. Uffizi Gallery has recorded over two million visitors per year, generating an income of 35.6 million euros in 2019. As to touristic attractions in Pont du Gard city, France, the same framework and metrics have been used to predict presences and computing causality.

The rest of the paper is structured as follows, see also Fig. 1. Section III provides details about data collection which refers to the blocks on the left in Fig. 1. In Section IV, the reputation analysis is provided with the aim of estimating reputation by exploiting Twitter based metrics in a more rapid way than waiting for Trip Advisor data. In this context, Sentiment Analysis (SA) has been used as a specific explainability (XAI) approach to compare domains of TripAdvisor reviews and Twitter data. Section V proposes a model for predicting the value of Progressive Mean of Reputation (PMR) using deep learning. Section VI reports the model for computing Multilingual Sentiment Analysis of Trip Advisor and Twitter data adopted in many other models, See Fig. 1, and that has turned out to be valid in Section IV. In Section VII.A, a predictive model to forecast the number of presences based on Twitter and other metrics is reported. Section VII.B provides a causality analysis of the number of presences which can be used to plan advertising and early warning. Conclusions are drawn in Section VIII.

Data collection

Computing online reputation and presence in the context of tourism attractions represents a complex challenge. By leveraging data-driven solutions, it becomes possible to harness information from heterogeneous data sources. As introduced in Section I.A different data sources have been used in literature, from simple tourism demand data to data coming from Social Media platforms. Regarding this work, the selected data sources are summarized in Table 2, which also provides details about the whole data collection process.

The selected data sources are related to three main categories:

- Data related to the number of tourists/visitors: for the Uffizi, they have been collected from official reports (https://uffizi-production-b8df82a1.s3.eu-central-1.amazonaws.com/production/attachments/1580731356430623-numeri-2019-definitivo-hd-v2.pdf); for Pont Du Gard any counting has been performed by IoT sensors and sold tickets: people counter and bike counter sending data into the

Table 2
Data sources and data collection tools.

Data source	Data collection
Uffizi reviews and scores from TA	TripAdvisor Uffizi's page
Uffizi number of visitors	Uffizi's reports
Tweets about Firenze and Uffizi	Twitter Vigilance [11]
Multilanguage SA annotated dataset about Tourism related tweets	Twitter Vigilance + Annotation tool [11]
Tweets about Pont Du Gard	Twitter Vigilance [11]
Pont Du Gard people counter sensors data, tickets, bike counting, etc.	Snap4City platform
Pont Du Gard bike counter sensors data	Snap4City platform
Weather data on Pont du Gard	WorldWeatherOnline

Snap4City platform. The platform provides APIs to gather information based on temporal windows.

- Data from touristic social media platforms: Trip Advisor regarding reviews and scores related to the Uffizi Gallery (https://www.tripadvisor.it/Attraction_Review-g187895-d191153-Reviews-Gallerie_Degli_Uffizi-Florence_Tuscany.html) and Florence, and Twitter data from Twitter Vigilance platform [11] with a multilanguage data set about tourism, as far as the Uffizi, Florence is concerned, while Pont Du Gard tweets were used for this location.
- Weather data of Pont Du Gard gathered from the World Weather Online (https://www.worldweatheronline.com/).

Attractions' reputation analysis

Reputation assessment of attractions in cultural heritage sites is a key topic when dealing with tourism management. To this end, the goal was to assess if the reputation extracted from Twitter and the one from Trip Advisor are somehow comparable / exchangeable. The two social media channels have also different user profiles, reaction times and volumes of posts/tweets as to Twitter, and stars/scores and comments as to Trip Advisor. TripAdvisor is a media platform allowing users to report a review or comment after they experienced a tourism related activity, and this may be published on the platform days after this activity took place. In Twitter, users rapidly share their content in the social network and tweets are typically posted during their visit and not after.

In this paper, the resulting data from the two platforms have been compared to discover if it is possible to estimate the Trip Advisor reputation from Twitter data, as well as the relationships with the number of visitors. To this end, the following data have been collected:

- **Reviews and scores from Trip Advisor** in the selected period, about 7300 regarding the Uffizi Gallery reputation. Trip Advisor scores are reported in a scale from 1 to 5 stars, where 3 is considered neutral, 4 and 5 positive and 1 and 2 negative.
- **Number of visitors/presences** at the Uffizi Gallery in the temporal window since January 2018 up to December 2019, two years of data. Data on presences have a monthly time granularity, i.e., 24 samples. Please note that Uffizi Gallery is the major attraction in Florence.
- **About** 2 M **of tweets/retweets** in the period since January 2018 up to December 2019, related to Firenze and Uffizi Gallery keywords, hashtags, citations, etc., acquired and processed in real-time through the Twitter Vigilance platform [11].

As a first step, since presences are provided monthly, data have been aligned to a monthly scale. As a second step, Trip Advisor reviews and Twitter tweets/retweets have been processed to estimate a number of metrics and a sentiment score. The approach used to estimate the sentiment score by exploiting sentiment analysis has been described in Section VI.

The summary of the metrics/features used in the dataset for assessing the relation among the Uffizi Gallery reputation provided by Trip Advisor, the number of presences/visitors in the Uffizi, and the Tweets related to Firenze, are reported in Table 3. As shown in the table, temporal features such as year and month have been also considered. Initially, only tweets/retweets containing the keyword "*Uffizi*" were considered, and it became very soon evident that the approach was too limited to perform predictions and assessment; thus, the choice was to include also tweets/retweets about Firenze. Among the metrics/features, the "volume" – i.e., the number of tweets/retweets – has been considered.

As to tweets/retweets, a sentiment analysis has been performed (described in more details in Section VI). The model described in Section VI provides, for each result, an output that reports a vector with the values $[p_{neg}, p_{neu}, p_{pos}]$, where p is the probability of negative, neutral and positive sentiment, respectively. For each temporal window, three sentiment analysis metrics have been estimated. They are the average of positive and negative scores, and the total sentiment score in the considered temporal window. The total score, *Tweets_Score*, is obtained by computing the difference of the two averaged *Tweets_Score_Pos* and *Tweets_Score_Neg* as follows:

$$Tweets\_Score\_Neg = \frac{\sum_{i=0}^{Tn} p_{negi}}{Tn}$$

$$Tweets\_Score\_Pos = \frac{\sum_{i=0}^{Tp} p_{posi}}{Tp}$$

$$Tweets\_Score = Tweets\_Score\_Pos - Tweets\_Score\_Neg$$

Where: *Tp* and *Tn* are the number of positive and negative tweets/retweets in the selected temporal window, respectively. Moreover, the

Table 3
Metrics/features for the reputation assessment: N for numbers, S for scores, and V for volume metrics.

Feature	Description	Range	VSN
Year	The year of the observation	{2018,2019}	N
Month	The month of the observation	{1, …, 12}	N
TripAdvisor_ Reputation	The mean score of the reviews from Trip Advisor about the Uffizi Gallery in the period	{4.335, 4.799}	S
Reviews_Total_ Volume	The volume of the Uffizi reviews acquired from Trip Advisor in the period	{175,…,471}	V
Negative_Reviews	The volume of negative reviews of the Uffizi from Trip Advisor in the period	{50, …, 154}	V
Neutral_Reviews	The volume of neutral reviews of the Uffizi from Trip Advisor in the period	{4, …, 38}	V
Positive_Reviews	The volume of positive reviews of the Uffizi from Trip Advisor in the period	{115, …, 290}	V
Tweets_Volume	The volume of tweets/retweets acquired about Firenze in the period	{48,862, …, 178,922}	V
Negative_Tweets	The volume of negative tweets/ retweets acquired about Firenze in the period	{18,450, …, 100,706}	V
Neutral_Tweets	The volume of neutral tweets/ retweets acquired about Firenze in the period	{3570, …, 17,692}	V
Positive_Tweets	The volume of positive tweets/ retweets acquired about Firenze in the period	{26,708, …, 755,888}	V
Tweets_Score_Neg	Average negative sentiment score in the period	{0.233, …, 0.493}	S
Tweets_Score_Pos	Average positive sentiment score in the period	{0.247, …, 0.546}	S
Tweets_Score	Tweets_Score_Positive minus Tweets_Score_Negative in the period	{−0.205, …, 0.314}	S
Uffizi_Presences	The number of visitors of the Uffizi Gallery in the period	{200,000, 480,000}	V

simple counting of the number of positive/negative tweets are considered as volume metrics.

Estimating reputation: regression analysis

The idea of estimating Trip Advisor reputation by means of other data is grounded on the fact that Trip Advisor score reputation is considered reliable by tourists during the selection of their next destination and it is accessible much later than the direct estimation of other considered variables, such as tweets. A linear regression analysis has been conducted to get information about the relation among the *Trip Advisor_reputation* of the Uffizi Gallery and the independent variables reported in Table 3. Multiple linear regression has aimed to understand regression capabilities of one or more features. Assuming the intercept to 0, N the number of features considered, the parametric formula is:

$$Y = \sum_{i=1}^{N} (\beta_i * X_i)$$

Where: β_i are coefficients to be determined. As a result of a large number of experiments performed while taking into account the variables of Table 3, two relevant cases/models have been identified, while many others have been considered as not or less significant. The results of the developed regression models are reported in Table 4. Given the limited amount of data samples regarding presences and Trip Advisor scores, 24 observations have been considered along the 24 months of data history.

In ***Case 1***, *TripAdvisor_reputation* has been found related to the *Uffizi_Presences* with a linear model. The scale factor has been of 1.2337e-05 and the *p-value* resulted much smaller than 0.05, stressing its significance: p-value =1.826867e-16. The estimation of *TripAdvisor_reputation* can be performed by using the *Uffizi_Presences* with a *MAE* of 0.8, and an R-squared (R^2) of 0.95. This means that reputation is positively influenced by the number of presences. This result has confirmed a linear model for the fact claimed in [5], namely that reputation is influenced by the touristic volume. More visitors you have and more likely is the increment of reputation, also because positive comments are easy to be posted on social than negative ones.

The R^2 has been calculated as:

Table 4
Linear Regression for estimating reputation.

Case 1			
Dependent Variable	TripAdvisor_Reputation		
independent variables	**Variable**	**Coefficient**	**p-value**
	Uffizi Presences	1.2337e-05	1.826867e-16
MAE	0.8011		
MAPE	17.254		
R-squared	0.950		
F-statistic	437.3		
Case 2			
Dependent Variable	TripAdvisor_Reputation		
independent variables	**Variable**	**Coefficient**	**p-value**
	Tweets_Volume	1.831e-06	2.517391e-02
	Tweets_Score_Neg	4.0488	7.660306e-06
	Tweets_Score_Pos	7.6894	9.744637e-12
MAE	0.0675		
MAPE	1.471		
R-squared	0,459		
F-statistic	19,280		

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n} (obs_i - pred_i)^2}{\sum_{i=1}^{n} (obs_i - \bar{y})^2} \right)$$

where: $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} obs_i$.

The MAE is calculated as:

$$MAE = \frac{\sum_{i=1}^{n} |obs_i - pred_i|}{n},$$

where:$obs_i = observationattimei$,

$pred_t = prediction\ at\ time\ t$,

$n$is the number of the values in the test set.

In ***Case 2***, *TripAdvisor_reputation* has been found related to the linear combination of *Tweets_Volume, Tweets_Score_Neg, Tweets_Score_Pos,* metrics of Table 3. All of them turned out to be significant estimators of reputation. The coefficient and p-values are reported in Table 4, variables with *p-value* larger than 0.05 have been excluded from the model due to lack of significance. The MAE in this case was quite low, R-Squared very high, as well as F-stat very high, thus remarking the high relevance of the model. It is therefore evident that the sentiment analysis of *tweets/retweets* can be a good estimator of *TripAdvisor_reputation.* According to this resulting model, decision makers can obtain Trip Advisor reputation, as soon as Twitter data are accessible. In Section VII a causality analysis has been reported.

Moreover, at least in the period where the assessment took place, no conditions of overcrowding have been found [43,44]: if that had occurred, the number of people would have affected the service quality of the attraction, thus receiving negative reputation scores [41]. Due to the low temporal resolution of 1 month, it is likely that any negative effect of over tourism was limited to some daytime slots or days in the month, whereas that effect disappears in monthly values.

In ***Case 3***, the usage of machine learning approach has been tested on the same data. The monthly time granularity of the Uffizi presences limited the possibility of using deep learning models to estimate reputation with only 24 samples. Ensemble learning techniques could be used to model cases with limited datasets as in [45]. Ramdani et al. in [46] stated that the XGBoost could be the solution, when dealing with small datasets. Zou et al. in [47] has used cross validation to handle the modeling of small dataset with XGBoost. Given these findings, a XGBoost model has been developed to assess reputation, on the basis of the features reported in Table 3 and aiming to understand which could be the most informative features related to Trip Advisor online reputation. As suggested in [47] cross validation has been used to fit this model, particularly in the configuration leave-on-out. The metrics evaluated on the set of the left-out folds report a MAE of 0.072, an RMSE of 0.097, MAPE = 1.568 and a R2 of 0.291.

The feature importance extracted from each cross-validation step has been summed up to obtain the overall feature importance as described in Fig. 2. The top 5 most important features reported in the plot are the

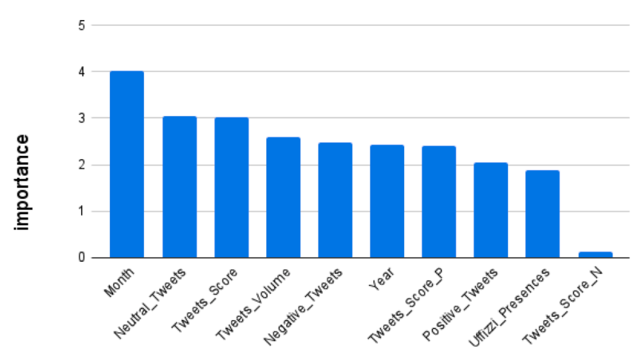

Fig. 2. Feature importance obtained on cross-validated XGBoost model.

Month and 4 other features from Twitter data in terms of volumes and the above defined scores. The interesting finding is that the features extracted by the considered social network turned out to be more important than the Uffizi Gallery presences.

Reputation analysis using explainable AI

Case 2 regression of previous section, as well as results on the feature importance obtained by the XGBoost model (**Case 3**), demonstrated that TripAdvisor reputation score is related to TW score metrics provided by sentiment analysis. In this section, it has been analyzed if a similar relationship is present also between Trip Advisor *reputation reviews* and *tweets/retweets with positive and/or negative sentiment*. To this end, XAI analysis has been performed for both Trip Advisor and Twitter, specifically on *reviews* and *tweets/retweets* within the two years period.

In order to explain the developed Transformer Sentiment Analysis model (whose details are reported in **Section VI**), local interpretable model-agnostic explanations (LIME) [42] has been chosen as XAI approach, on top of which a new representation model has been developed, described as follows.

For each Twitter tweet/retweet and TripAdvisor review, 10 most important keys/features have been extracted by the Sentiment Analysis explainability model of the predicted class computed as reported in **Section VI**. If such keys positively contributed to the predicted output of the Tweet/Retweet t, the value ($LimeValueK_t$) assigned by the explainability library was positive, and negative otherwise. The next step has been to process each tweet/retweet and review and save the corresponding sentiment analysis XAI results. 10 most important features of reviews and tweets/retweets have been ranked on the account of their $Importance_K$ metric, defined as the sum of the absolute value of key values $LimeValueK_t$ for every Tweet/Retweet in the dataset. Formally:

$$Importance_K = \sum_{t=0}^{T} |LimeValueK_t|$$

Where T is the total number of Tweets/Retweets in the dataset and K ranges for the top 10 keys evaluated in the explainability process described above.

For such top 10 words of both sources (reviews and tweets/retweets) a plot has been produced displaying the distribution of the values assigned by the explanation in the resulting range. This procedure was carried out for positive and negative classes, and results are reported in Fig. 3 for TripAdvisor reviews and Fig. 4 for Twitter tweets/retweets.

In particular, the plots of Figs. 3 and 4 (which are new representations for global explainability on the basis of LIME data, see **Section VI**) represent the distributions of the explained Sentiment Analysis computed by Transformer highlighting the most relevant words/features reported on the left side of the plots. The order of the words has been considered according to the relevant ranking of the words in the total set determined by the explanation analysis conducted with LIME. In graphs of Figs. 3 and 4 the *x-axis* reports the resulting contribution towards the predicted class per each word.

For example, in the negative plot of Fig. 3(i), a peak on $x = 0.3$ for "crazy" means that the word in the sentences has contributed positively quite often (the *y-axis* describes the frequency by which that score has been registered) towards the predicted class to the estimation of negative sentiment. The same word does not have contributed negatively, as evident from the lack of negative data in the diagram. On the contrary, "opera" has contributed negatively quite often in a moderate way, while the positive contribution is present, though rarely. The height of the *y-axis* on the plot to a specific value represents how many occurrences of that word resulted in a specific contribution as a distribution histogram: the higher the position of the point, the greater the number of times the word has occurred in the considered groups of most relevant features. The total number of negative and positive votes, the area under the curves, could be regarded as the amount of positive/negative impact.

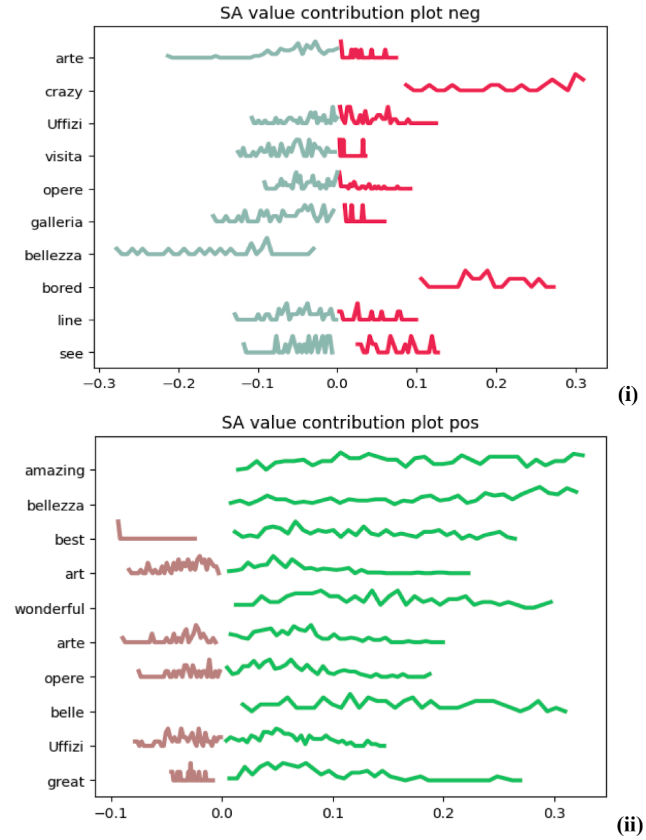

Fig. 3. – Contribution plot of the 10 most important words in the sentiment analysis computation for TripAdvisor Reviews: (i) Negative (red lines), (ii) Positive (green lines) to simplify the reading. The x-axis indicates the resulting contribution of each word towards the predicted class reflecting both positive and negative influences, highlighted with a contrasting color line in comparison to the green for positive and red for negative. The y-axis represents the frequency of occurrence of each contribution value.

Considering the results for the positive class (see Fig. 3(ii)), the top 3 most important features regarding the tweets are "amazing" "bellezza" and "best", whose explainable values almost always did contribute positively to the positive predicted SA class. This polarity is also evident for the top 3 most important features regarding TripAdvisor negative reviews (see Fig. 4(i)**)** where "ucciso" (killed), "omicidio" (murder), "spara" (shoot), have contributed almost always towards the negative predicted SA class. There are also disputed aspects as "senegalese" (coming from Senegal) and "Renzi" (a politician) which do not provide a defined orientation, and drove a similar amount of positive and negative comments.

From a first-glance analysis on the words identified in the XAI reported on Figs. 3 and 4, it is quite evident that the distribution of words in Trip Advisor and Twitter seems to be almost disjoined. For this reason, a deeper study varying the number of most important words from the XAI of the estimated SA from Trip Advisor and Twitter has been performed, with the aim of identifying at least 10 matching words.

Most of the words reported in Figs. 3 and 4 are in Italian. This is due to the fact that in both cases, the original data regarding Trip Advisor and Twitter texts are in Italian, as reported in the distribution pie of Fig. 5.

Results shown in Fig. 6 explain that: in order to get 10 matches from the two XAI models of Sentiment Analysis for both Trip Advisor and Twitter data, about 120 features/words have to be taken into account. As demonstrated in Fig. 5, the distribution of languages used in Trip Advisor and Twitter is very similar. This analysis has demonstrated that the word "space" used by Trip Advisor and Twitter users seemed to

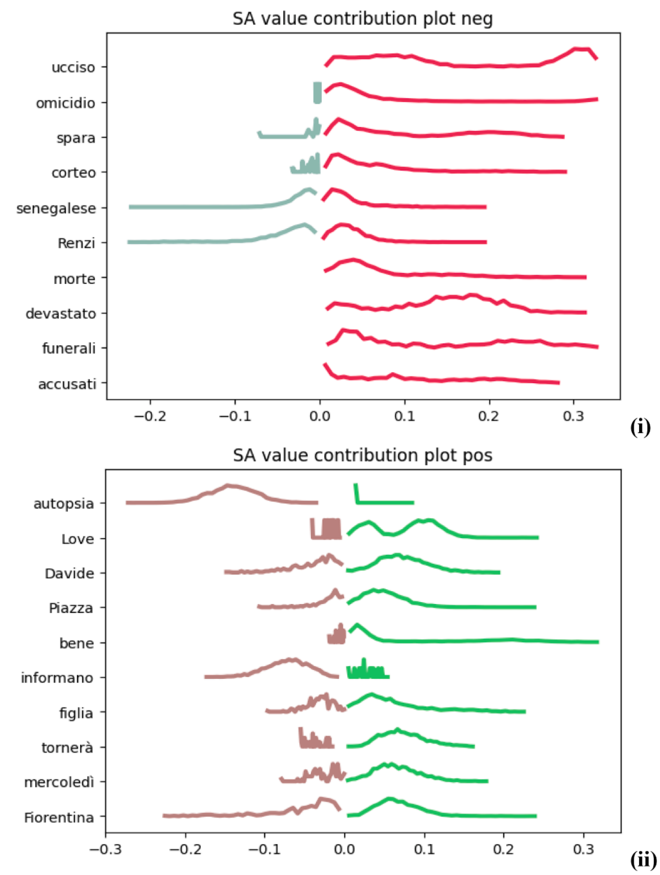

Fig. 4. – Contribution plot of the 10 most important words in the sentiment analysis computation for Twitter data: (i) Negative (red lines), (ii) Positive (green lines) to simplify the reading.

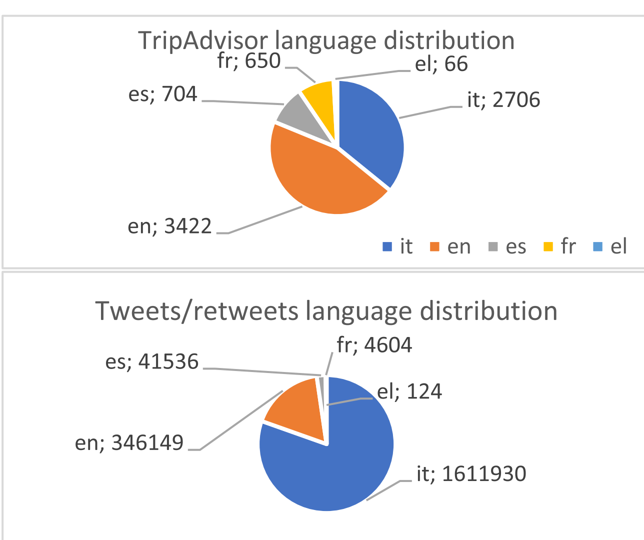

Fig. 5. – Language distribution for both TripAdvisor reviews and Twitter tweets/retweets.

belong to quite different slangs, whereas it produced the same results in terms of reputation. This fact may depend on the differences of user profiles in the two platforms [48,49].

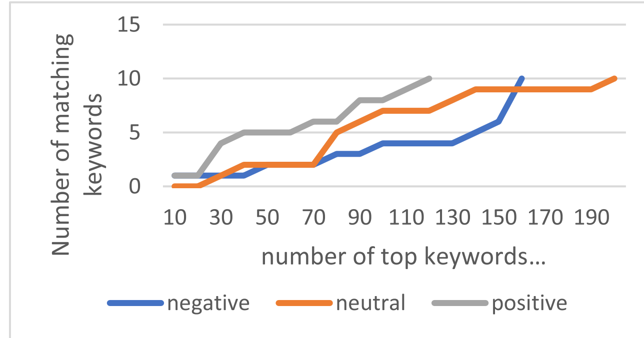

Fig. 6. – Number of matching words (y) plot (in both reviews and tweets) depending on the total number of most important words XAI of the sentiment model.

Prediction of reputation

The analysis performed on Section IV Case 2 has suggested that metrics computed from TW data could be a promising solution for assessing any reputation/appreciation customers/visitors posted on Trip Advisor in terms of stars. The number of reviews posted on Trip Advisor about the Uffizi Gallery is around 7300 within the two years period. This means a mean of 10 reviews per day, with a minimum of 0 reviews and a maximum of 32 reviews per day. On the other hand, the volume of tweets/retweets per day has registered a minimum of 92 and a maximum of 21,254. This large amount of data determined a large volume of information which daily can be processed for computing volume and sentiment analysis metrics, as reported in Table 3. In Fig. 7, the trend of monthly average Trip Advisor stars/scores is reported for the two years period of observation. It is self-evident that the most appreciated periods are from December to April.

From this kind of a-posteriori analysis it is possible to understand if the trend is improving, while it is difficult to get the trend day by day. A different analysis can be obtained by performing an assessment day by day of the progressive mean reputation/appreciation, PMR. The PMR is the mean of the scores in terms of stars registered from the beginning of the year up to the day under consideration, and it is based on the computation of the Daily Mean Reputation, DMR, on the basis of the number of reviews obtained in day *d*, R_d:

$$DMR_d = \sum_{i=0}^{R_d} ReviewReputation\_i$$

$$PMR_t = \frac{\sum_{d=0}^{t} DMR_d}{t}$$

Where: t is the day of the year considered $t=\{1,\ldots,365\}$. The PMR for the years 2018 and 2019 are reported in Fig. 8.

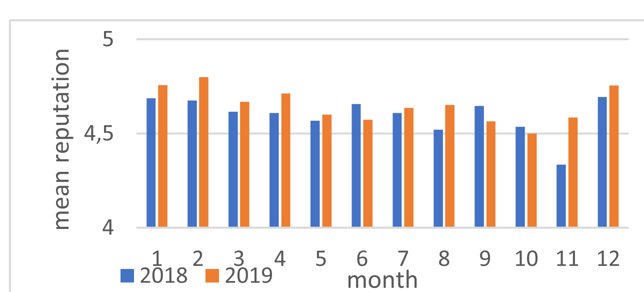

Fig. 7. – Monthly average reputation/appreciation from Trip Advisor stars.

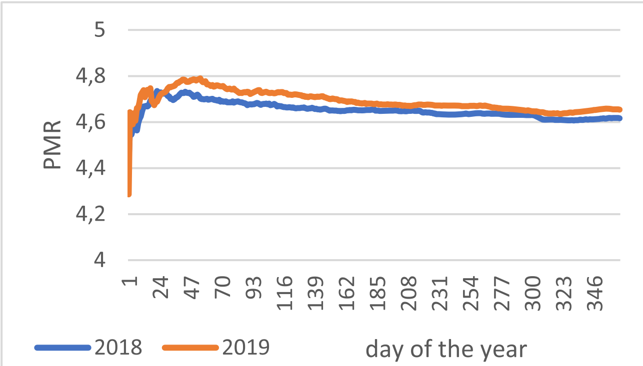

Fig. 8. – Progressive Mean of Reputation, PMR, from TA.

Prediction of progressive mean reputation

The daily estimation / prediction according to **Case 2 of Section IV** is an improvement with respect to getting the score directly from Trip Advisor, in which the reviews arrive 15–30 days later the actual visits of the tourists. Moreover, according to the trend of reputation/appreciation, a relevant result could be to produce a *prediction of the reputation trend at least a month in advance*. Since the data may come every day from Twitter, the daily prediction of PMR would be feasible.

The set of data considered for the predictions has been improved from those of Table 3, in order to make them suitable for the development of the predictive models. The first step has been to set a daily time resolution in the same time window from January 2018 to December 2019. From the temporal features: (i) the Year metric has been removed, (ii) a feature has been added coding if the day is a working day or a weekend (Saturday or Sunday), (iii) the feature coding the month of the year has been maintained, (iv) a feature to code the day of the week has been added. Day of the week and month features have been coded to describe the cyclic model by using a cosine transformation, such as:

$$DayofTheWeeksin(x) = \sin\left(\frac{2\pi x}{7}\right)$$

$$DayofTheWeekcos(x) = \cos\left(\frac{2\pi x}{7}\right)$$

$$Monthsin(x) = \sin\left(\frac{2\pi x}{12}\right)$$

$$Monthcos(x) = \cos\left(\frac{2\pi x}{12}\right)$$

Regarding the information retrieved from Twitter, the metrics used have been Tweets_Volume, Negative_Tweets, Neutral_Tweets, Positive_Tweets, Tweets_Score_Neg, Tweets_Score_Pos, Tweets_Score. Regarding the Trip Advisor metrics, the Reviews_Total_Volume has been used in addition to the target variable TripAdvisor_Reputation.

On the basis of the above data, a number of machine and deep learning models have been developed and compared: XGBoost [19] regarding the forecasting methods based on ensemble learning, Temporal Convolutional Network for the Deep Learning techniques (TCN) [20], and Temporal Fusion Transformer Networks (TFT) [21].

Regarding the training process, the same strategy has been applied for all the 3 models, i.e., an Early Stop on the validation MAE, patience set at 10, performing a hyperparameters optimization procedure with 100 trials in the space defined, described later for each model.

Regarding the **XGBoost** the hyperparameters optimized have been the maximum depth in the range {3, …, 10}, and the learning-rate in the range of (0.001, 0.08).

The **TCN** has been optimized based on a total of 30 timesteps and enabling the possibility to use 1 Temporal Convolutional, (TC) layer or stacking up to 4 TC layers {1, 2, 3, 4}. The other hyperparameters optimized have been the learning-rate in the range {0.001, …, 0.01}, the kernel size among {3, 6, 12, 24}, the number of filters to use in the convolutional layers among {16, 32, 64, 128}, the number of stacks of residual block to use in {1,2,3,4}, the dropout rate in the interval {0,0.1, …,0.5}.

The **TFT** has been developed assuming a max encoder length of 30 and clipping the gradient to prevent the divergence of these; the range of optimization for this hyperparameter has been {0.01, …, 1}. The other hyperparameters optimized have been the learning-rate in the range of {0.1, …, 0.8}, the dropout in {0., 0.1, …, 0.8}, the network size defined by the hidden size in {8,10, …, 128} and the continuous hidden size equal to half the hidden size. The last parameter considered has been the number of attention heads in the range of {1,2,3,4}.

The above-mentioned models were evaluated in terms of statistical measures such as R^2, MAE, Root Mean Squared Error (RMSE), and MAPE. The RMSE has been used to compare the models and to choose the best solution for the short-term predictions of the PMR. The RMSE is calculated as the root square of the MSE:

$$MSE = \frac{\sum_{i=1}^{n}(obs_i - pred_i)^2}{n}$$

$$RMSE = \sqrt{MSE}$$

The MAPE is calculated as follows:

$$MAPE = 100\,\frac{\sum_{i=1}^{n}\left|\frac{obs_i - pred_i}{obs_i}\right|}{n}$$

where:

$$obs_i = observation\ at\ time\ i,$$

$$pred_t = prediction\ at\ time\ t,$$

*n*is the number of the values in the test set.

The split of data for training, validation and test has been performed by considering that the data availability covered 2 years, and it has been set to: 0–45% as training set, 45%−55% as validation set and 55%−100% as test set. This allowed to perform a validation over the two years since there was not a third year to perform an additional validation.

The different solutions were compared on the basis of the defined metrics as reported in Table 5. The mean value of the reputation in the test set used was 4.686, which is useful to understand the offset due to the fact that it has been used the MAE, and MSE which depend on the magnitude of the values.

The best model resulted to be the TFT, with the best hyperparameters such as: 2 attention heads, dropout of 0.347, gradients clipped at 0.995, 100 hidden size and 50 continuous hidden size, learning-rate of 0.0432. The best approach resulted to be the TFT model. It has one TFT layer and the other hyperparameters have been: dropout of 0.276, kernel size of 12, learning-rate of 0.0018, 32 filters in the convolutional layer and, 2 stacks of residual block. The hyperparameters for the XGBoost resulted to be: a learning-rate of 0.143 and a maximum depth of 7.

The trend of predictions with respect to the actual values of the progressive average of the TripAdvisor reputation is displayed in Fig. 9. Please note that the graph limited the reputation values from 4.63 to

Table 5
Comparison of models for predicting PMR: 1 hour forecast.

Model	MAE	RMSE	R2	MAPE
XGBoost	0.0375	0.0409	0.892	7.9%
TCN	0.0368	0.0384	0.942	7.8%
TFT	**0.0041**	**0.0056**	**0.973**	**0.8%**

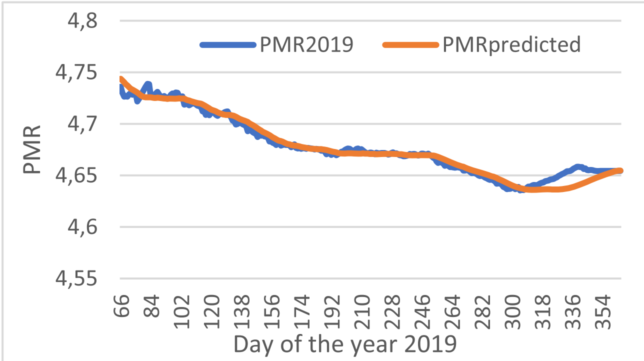

Fig. 9. – Computed predictions of PMR by using TFT with respect to the actual reputation progressive averaged score along the days of the year 2019.

4.75 to show in more detail the predictive precision. Please note that comparing Figs. 7 and 9, the reputation seems to be strongly related to the number of visitors, and the predictive PMR confirms this fact.

Feature importance for prediction reputation models

A further analysis has been made about the feature importance on the evaluated models regarding the prediction of PMR. The XGBoost library used provides the possibility of extracting the feature importance after the training of the model. This functionality is also provided by the TFT whose explainability is provided for the most important encoder variables as well as for the decoder part of the architecture. However, the XAI library LIME [42] has been used to obtain the explainability from the TCN. The model works on 13 features and a temporal window of 30 timesteps. Thus, in order to obtain a global feature importance, a specific procedure has been adopted as described in the following.

Considering all the samples in the test set, the LIME library associates an explainability value *LimeValueF_t*, where *F* is the considered feature and *t* is the timestep. For each instance of the test set, 13×30 values are cumulated computing the total value of importance for each feature as:

$$Importance\ (F) = \sum_{t=1}^{30} |LimeValue\ (F_t)|$$

The results of the importance of each feature are sorted to obtain the final ranking as reported in Table 6 for the first 3 of each model, in which several similarities can be observed. The results on the top 3 most important features slightly differs from model to model.

According to Table 5, the best architecture for PMR computing reputations is TFT and the explainability confirmed the features relevance of the model estimated in the Case 2 of multilinear regression: number of positive and negative tweets coming from Sentiment Analysis of Section VI. The Twitter post Sentiment Analysis metrics are the most important considering all the 3 models developed with 6 positions in the top 3. It should be noticed that also the seasonal information regarding the month (with its coding monthsin, monthcos) is very relevant, confirming the results of Fig. 7.

Computing sentiment analysis

According to the previous sections, social media analysis can provide useful information for tourism analysis: computation of the reputation of attractions and locations, prediction of the averaged value of reputation, etc. There is a large variety among the social media channels in terms of reaction time of their users, as compared in the previous section Twitter is much faster than Trip Advisor, and both of them produce text which are tweets/retweets and reviews, respectively. Computing sentiment analysis can be of extreme importance to understand if the content that is generated on the social network is leading the users towards a positive discussion or even more dissatisfaction.

The platform for collecting tweets has been Twitter Vigilance [11], which is grounded on the definition of *Channel* which consists of a set of simple and complex queries performed on Twitter platform. *Complex Channels* consist of tens of queries according to Twitter API syntax combining **keywords, hashtags**, etc. The TripAdvisor reviews have been manually collected.

For sentiment analysis, SA, the platform has been extended with an annotation tool for multilingual SA, enabling mother tongue experts to annotate the text relevant to the tourism with a SA. To this end, the user interface reported in Fig. 10 has been used to annotate about 5000 tweets per language: English, Italian, Greek, French, Spanish. The categories for the SA were negative, neutral, positive and skip if the mother tongue expert did not want to express the sentiment on particular texts for different reasons: if they were uncertain or if the context was not tourism-related.

A multilingual dataset has been made of a total of 23,665 annotated texts: 4729 English tweets, 5030 Italian, 4863 Greek, 4257 French and 4786 Spanish. The distribution of the sentiment classes is reported in Fig. 11, from which it is evident that the distribution of classified text differs per language. The Spanish and Greek annotated texts have similar distribution with the prevalent classes negative and neutral. The French and English texts prevalent class is the neutral with more negative than positive annotated tweets. The Italian annotated texts have balanced number of negative and positive tweets with minor number of neutral texts. This unbalanced distribution is not ideal and could create biases in the SA model. Moreover, using only a subset of the dataset will mean having a dataset of only 7.590 texts because the minimum class is the neutral Italian with 506 texts.

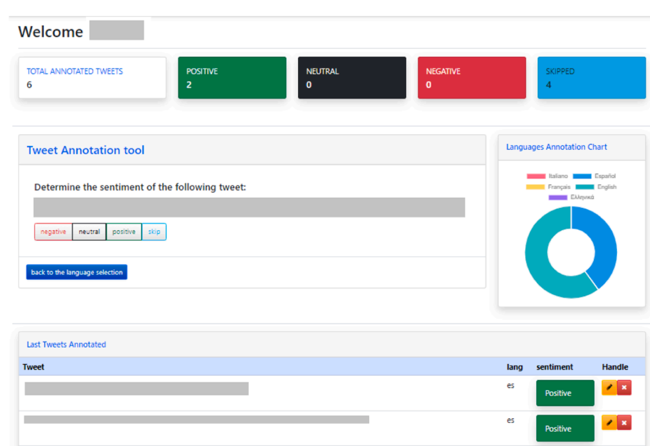

Fig. 10. Twitter Vigilance annotation tool user interface for texts.

Table 6
Feature importance on the AI models considered.

Imporance	XGBOOST	TCN	TFT encoder	TFT decoder
1	Monthsin	Tweets_Score	Positive_Tweets	Monthsin
2	Positive_Tweets	DayofTheWeeksin	Negative_Tweets	DayofTheWeeksin
3	Reviews_Total_Volume	Tweets_Score_Neg	Tweets_Score	Monthcos

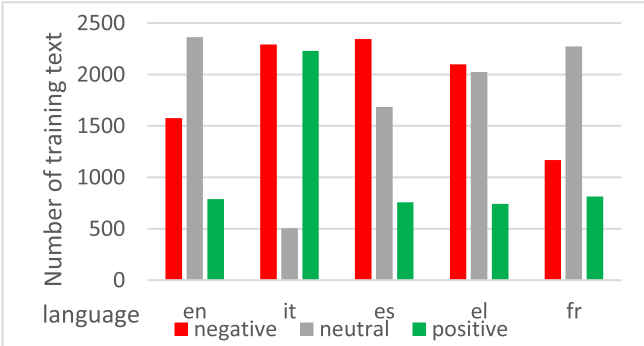

Fig. 11. Distribution of number of training text per language.

Sentiment analysis AI model development

In this section, the model developed for Sentiment Analysis computation used in **Sections IV and V** is described.

The current state-of-the-art for computing Sentiment Analysis is based on Transformers architectures. In most cases, the classification of sentiments is much more complex than positive/negative. On the other hand, in this case, Sentiment Analysis has been used for assessing the reputation of tourism attractions, which is in most cases rated with a set of stars, from negative to positive scores.

In the literature, there are some specialized BERT for tourism [50], while they are not multilanguage. Therefore, the proposed sentiment analysis model has been developed by performing a ***fine-tuning*** of a multilingual Bert Transformer model [51]. More specifically, the above-mentioned data set of annotated tweets has been split in 80% for training, 10% for validation and 10% as a test set. In particular, negative, neutral and positive classes were considered as output classes to obtain a trained model capable to distinguish the sentiment of the text. The fine-tuning process was implemented with Early-Stopping strategy with patience of 15 epochs to prevent the model to overfit. The batch size was 16 and the learning-rate of the Adam optimizer has been set at 0.00005.

Considering as class values “Negative”, “Neutral”, “Positive”, recalling the definitions of true positive (TP) as an outcome where the model correctly predicts the considered class, false positive (FP) as an outcome where the model incorrectly predicts the considered class, and false negative (FN) as an outcome where the model incorrectly predicts the not considered class, the following metrics have been computed per each class:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Also the accuracy metric has been reported considering all the “Negative”, “Neutral”, “Positive” classes with *Support* as the sample size of the class:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The results on the test set are reported in Table 7 and in the Fig. 12 in the form of a confusion matrix. The overall accuracy of the model over all the sentiment classes is 84%.

Table 7
Assessment Metrics for the Sentiment Analysis Model.

Language	Precision	Recall	F1-score	Support	
all	Negative	0.84	0.87	0.86	921
	Neutral	0.82	0.90	0.86	889
	Positive	0.88	0.69	0.77	558
	accuracy	0.84	2368		
it	Negative	0.86	0.88	0.87	215
	Neutral	0.83	0.68	0.75	57
	Positive	0.85	0.87	0.86	231
	accuracy	0.85	503		
en	Negative	0.80	0.79	0.80	166
	Neutral	0.81	0.83	0.82	235
	Positive	0.63	0.61	0.62	72
	accuracy	0.78	473		
es	Negative	0.83	0.92	0.87	221
	Neutral	0.85	0.74	0.79	176
	Positive	0.63	0.61	0.62	82
	accuracy	0.80	479		
fr	Negative	0.75	0.82	0.78	110
	Neutral	0.90	0.79	0.84	226
	Positive	0.58	0.68	0.62	90
	accuracy	0.77	426		
el	Negative	0.83	0.84	0.84	215
	Neutral	0.77	0.75	0.75	200
	Positive	0.51	0.54	0.54	72
	accuracy	0.75	487		

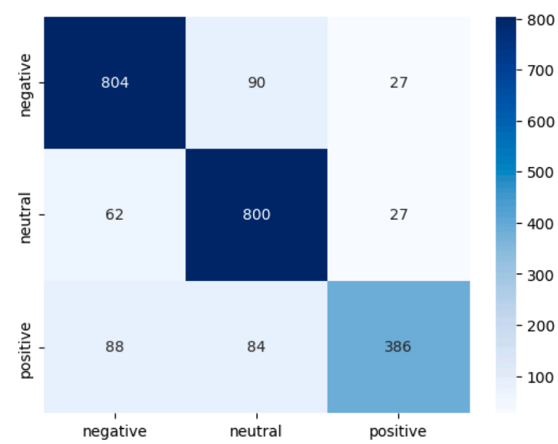

Fig. 12. – Confusion Matrix of the Sentiment Analysis model developed on the text.

Predicting presences

Overcrowded situations can cause dissatisfaction in tourist sites and attractions [43,44,53]. Knowing in advance the number of visitors/presences can be an important tool for operators, municipalities and also for visitors to improve the overall quality of experience. *This section aims to demonstrate that the Twitter metrics (volume and SA metrics) can be profitably used for predicting the number of visitors.* This case study can be regarded as a confirmation of the validity of Twitter metrics for predicting the number of presences. A distant and similar work was presented for predicting the number of attendees at the TV shows in [11]. In this case, the solution aims at predicting number of presences in specific locations and a number of additional features have been added as explained in the following.

For this purpose, a different data set of the presences has been used, since the data regarding presences at the Uffizi Gallery provided only monthly data. In fact, the case study is based on the number of presences on Pont-Du-Gard cultural attraction. The Twitter data have been gathered from Twitter Vigilance using a set of keys tuned on Pont du Gard attractions. The sensors data have been gathered from Snap4City platform and infrastructure. All the acquired data covered the temporal windows of 6 months from 2021 to 06–30 up to 2021–12–31, in the context of Herit-Data project. **The time granularity is hourly**, and the location provided also the counting of the number of bikes retrieved by 10 sensors, in addition to the 14 people counting sensors. The sensors are located in tourist functional points in order to monitor the total tourist audience situation, as shown in Fig. 13, in which the position of the nearest 3 people counting sensors closer to the attraction site (the bridge, "*pont*" in French) are reported.

A summary of the metrics used in the dataset for the prediction of presences is reported in Table 8, in which the measurements from people and bike counting sensors have been cumulated. In addition, there were also considered temporal features, weather features and the TW data volume and SA metrics computed by the model reported in **Section V.**

Modeling predicting presences tool

In order to create a reliable solution to compute predictions of people presences, three state-of-the-art models, i.e., XGBoost, TCN and TFT have been compared. The models are the same as the ones adopted in Section V for predicting reputation, while they have been applied on different data, with a different time coding, and for a different purpose. On the other hand, similar configurations have been used, except for the number of timesteps for the TCN and the max encoder length for the TFT that has been set to 24 h (1 day) tuned for this problem. The dataset reported in Table 8 has been tuned in order to be suitable for the development of the predictive models. In this case the coding for sine/cosine has been applied also on Hour metric, with period of 24 (hours):

$$Hoursin(x) = \sin\left(\frac{2\pi x}{24}\right)$$

$$Hourcos(x) = \cos\left(\frac{2\pi x}{24}\right)$$

The training process has been similarly applied: same strategy for all the 3 models developed with an Early-Stopping on the validation MAE with a patience set at 10 and a hyperparameters optimization procedure with 100 trials in the space defined. The dataset has been split in training, validation and test sets. From the whole dataset 5 days have been randomly extracted (each of which with 24 sequential hours) for the validation set and 5 for the test set, which have been removed from the training set. The machine learning solutions were compared in terms of MAE, RMSE, R2. The results are reported in Table 9. The target was to predict the total number of people arriving the next day (24 h in advance), in the next week (168 h), and in the next month (720 hour). The mean number of people in the test set for the total people in the next 24 h was 1961.169, for the week forecast 11,293.737 and for the monthly forecast 52,919.088.

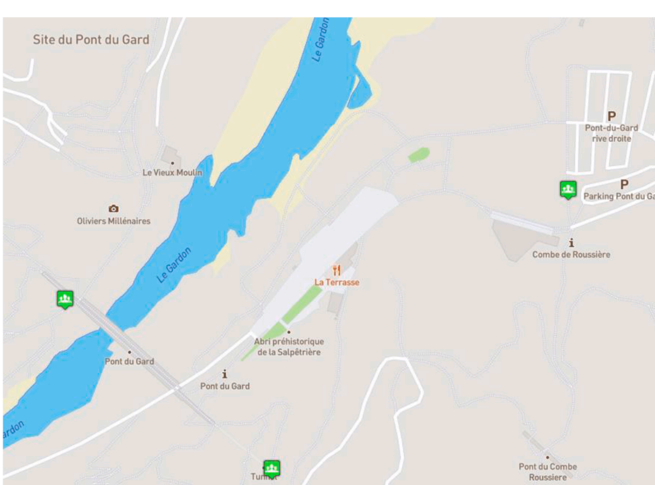

Fig. 13. – Locations of the most relevant sensors at Pont-Du-Gard.

Table 8
Metrics for predicting presences in Pont du Gard. All metrics are computed hour by hour.

Feature	Description	Range
peopleCounted	The number of visitors during the hour of observation	{0, …, 2021}
bikesCounted	The number of bikes during the hour of observation	{0, …, 210}
Hour	The hour of the observation	{1,…,24}
DayofTheWeek	The day of the week	{1, …,7}
IsWeekEndDay	If the day was Saturday or Sunday 1, 0 otherwise.	{0, …, 1}
Month	The month of observation	{1, …,12}
Temperature	The mean temperature of the hour of the observation in °C	{0, …,42}
Humidity	The mean humidity of the hour of the observation in%	{19, …, 98}
Precipitation	The millimeter of rain during the hour of observation	{0, …, 3.1}
Pressure	The mean air pressure in millibar (mb) of the hour of observation	{998, …, 1032}
WindSpeed	The mean value of the wind of the hour of observation in km/h	{0, …, 50}
Tweets_Volume	The number of Tweets in the hour of observation	{95, …, 19,475}
Negative_Tweets	The number of negative tweets/retweets in the hour of observation	{33, …, 8917}
Neutral_Tweets	The number of neutral tweets/retweets in the hour of observation	{60, …, 17,710}
Positive_Tweets	The number of positive tweets/retweets in the hour of observation	{2, …, 6299}
Tweets_Score_Neg	Average sentiment score negative in the hour	{0.000, …, 0.481}
Tweets_Score_Pos	Average sentiment score positive in the hour	{0.000, …, 0.501}
Tweets_Score	Tweets_Score_Positive minus Tweets_Score_Negative in the hour	{ −0.257, …, 0.271}

Table 9
Model comparison for predicting presences.

Hours	Model	MAE	RMSE	R2	MAPE
24 (1day)	XGBoost	83.430	153.644	0.977	5.7
	TCN	253.694	332.533	0.925	20.8
	TFT	43.536	68.261	0.995	2.9
168 (Week)	XGBoost	526.513	3021.86	0.756	4.3
	TCN	434.557	571.067	0.987	4.1
	TFT	275.542	365.466	0.994	2.8
720 (Month)	XGBoost	1193.39	2996.06	0.966	2.9
	TCN	8012.07	8428.48	0.769	18.4
	TFT	676.542	925.295	0.997	1.3

The best resulting model architecture has been the TFT for all the 3 forecasting temporal windows. The results reported a MAPE of 2.9 % for the number of daily presences, 2.8 % for the weekly case and 1.3 % for the monthly number of presences.

The hyperparameters for the TFT for the three temporal targets are reported in Table 10.

Regarding the comparison with the state-of-the-art, we can assume

Table 10
Hyperparameter of the best resulting TFT architectures for the forecasting targets next day, 7 days, 30 days.

	24 h	168 h	720 h
Attention heads	3	2	4
Dropout	0.031	0.193	0.119
Gradient clip value	0.149	0.3732	0.932
Hidden size	14	8	114
Continuous hidden size	7	4	57
Learning Rate	0.0219	0.112	0.157

that the number of tourist arrivals and presences could be the same feature kind. Thus, our target is equivalent to that in [29] regarding the next week prediction, and to the work [27] for the temporal forecasting window of 1 month. The most suitable metric of comparison is the MAPE, since it is a percentage value independent on the scale. According to the results provided in terms of MAPE in Table 9, the TFT with 2.8% outperformed both the RF of [29] that achieved a MAPE of 4.09% for 1-week forecast, and the LSTM of [27] with a MAPE of 9.2%, for 1-month forecast. The trends of the resulting predictions with respect to the actual values of presences are displayed in Fig. 14.

Causality analysis

In order to perform a causal analysis, a Granger causality analysis [52] has been conducted for the temporal shifts from 1 day, 2 days, …, 30 days regarding the time-series data reported in Table 8. On the basis of the Granger causality analysis, it cannot be rejected the null hypothesis for the variable X Granger (features) causes the variable Y (prediction) for the selected lag if the p-value is greater than the threshold defined at 0.05. The test has been generated for the temporal targets from 1 to 30 days by considering the daily *peopleCounted* (24 h) as target, and the time-series of features from Table 8 as causes. The features for which the test identified a causality for some of the time shifts are reported in Fig. 15.

Other variables tested which do not produce significant results have been: bikesCounted, Temperature, pressure, Humidity, WindSpeed, Tweets_Volume, Negative_Tweets, Neutral_Tweets.

The Granger causality revealed a strong causality for the temporal shifts of 22,…25 and in these cases it is not possible to reject the null hypothesis that the number of positive tweets and the Tweets score metrics Granger cause the daily *peopleCounted*. This result is important because it shows the importance that the information acquired through the SA on the data retrieved from TW on the monitored channel could be used in the case study of Pont-Du-Gard for: (i) the development of advertising to get presences, and to (ii) predict the presences the day before and all the older data regarding the scores. The weather conditions influence the visitor with a range of at least 3.5 days, a strong effect is produced by activities in the previous week.

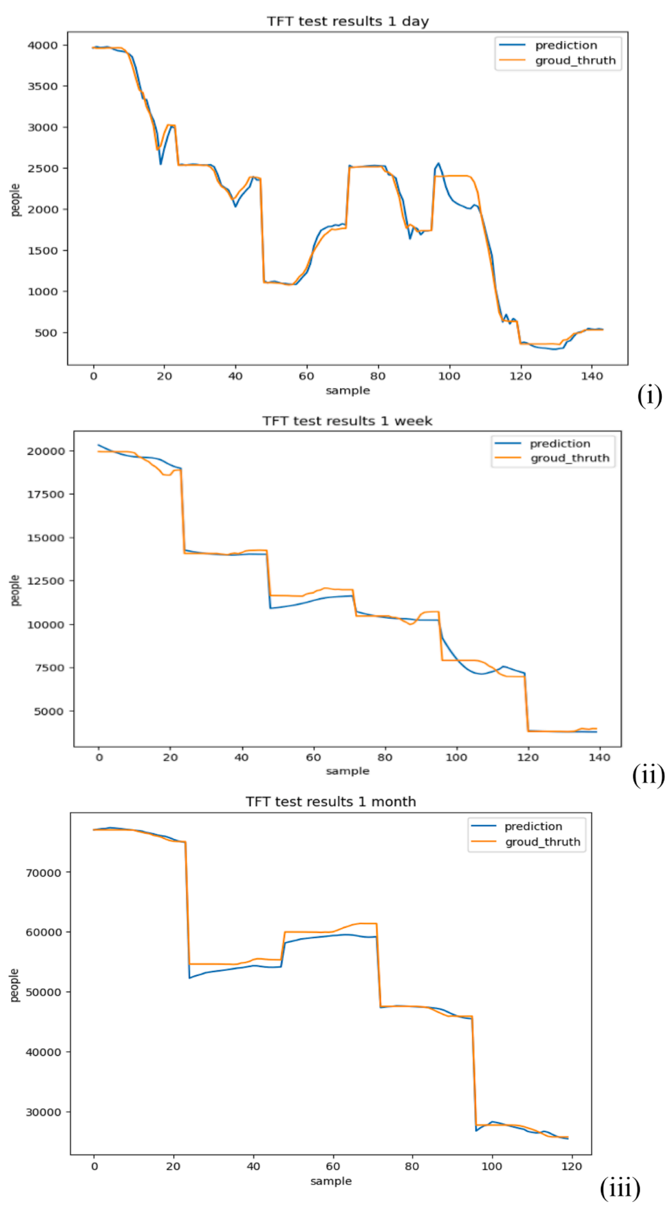

Fig. 14. – Trends presences predictions by using TFT wrt the actual number of presences. (i) next 24 h; (ii) next 168 h; (iii) next 720 h;.

Conclusions

The online opinion of tourism attractions is gaining more and more importance due to the digital innovation. People shares on social media opinions and experience, and these data can be used to develop systems supporting decision makers of tourism services in the cities. The frameworks to analyze the touristic attractions are complex since they need to involve the analysis of multiple data sources. Decision makers seek frameworks capable of assessing the online reputation of attractions, forecasting the indicators, and predicting the number of presences in advance. The motivations behind are related to improve the services, improve the tourist experience and market strategies (which also implies to understand the cause effect phenomena from promotion on social media to the number of presences). To this end, we have proposed a number of techniques based on machine learning, deep learning, causality assessment, and XAI. For this reason, a relevant element has been the multilingual sentiment analysis model developed to extract the opinion of the tourist online content with a precision of 84 % over the 5 languages considered, that has been used to obtain a set of sentiment analysis metrics on Twitter data. This model, harnessed in conjunction with Twitter data and counting metrics, has enabled the derivation of a set of sentiment analysis metrics, affording deeper insights into online tourist content. The proposed model for analyzing the explainability of the tourism data resulted in useful information for the problem of the assessment and prediction of online reputation reported on the online

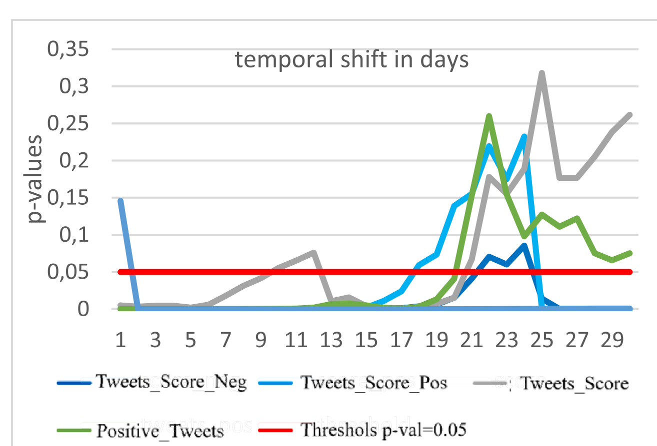

Fig. 15. – Granger causality for *peopleCounted*: p-values resulted from the causality analysis with the case reported metrics, in terms of temporal shift reported in hours (range of a week).

tourism platform Trip Advisor for the Uffizi gallery in Florence, Italy.

The research on the state-of-the-art forecasting architectures in the context of metrics related to smart tourism revealed a gap with the novel time-series forecasting architectures.

The proposed framework combines BERT sentiment analysis, forecasting AI architectures, and XAI techniques. For the prediction of the progressive reputation assessment the proposed TFT model achieved a R2 of 0.9727. This state-of-the-art architecture for the problem of time-series analysis resulted to be suitable also for the problem of forecasting reputation on Uffizi Museum, and the presences on Pont-Du-Gard, France, with the long-term temporal forecasting target of 1 day, 1 week, 1 month with a MAPE of respectively of 2.9 %, 2.8 % and 1.3 %. On the basis of the same data, a causality analysis has been performed to identify which are the drivers to be taken into account for very long terms prediction. Thus, we have verified that Twitter metrics provides some cause-effect with the number of presences for tourism attractions with a delay of about 3 weeks. This data can be used for planning advertising and preparing events.

CRediT authorship contribution statement

Enrico Collini: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Paolo Nesi:** Conceptualization, Methodology, Formal analysis, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Gianni Pantaleo:** Software, Validation, Writing – review & editing, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgement

Authors would like to thank the Herit-Data Interreg project of European Commission and partners. Snap4City (https://www.snap4city.org) is an open technology and research by DISIT Lab, University of Florence, Italy.

References

[1] A. Inversini, L. Cantoni, Cultural destination usability: the case of visit bath, in: W. Hopken, U. Gretzel, R. Law (Eds.), Information and Communication Technologies in Tourism 2009 – Proceedings of the International Conference in Amsterdam, Netherland, Springer. iProspect, Wien, 2009, pp. 319–331.
[2] S.W. Litvin, R.E. Goldsmith, B. Pan, Electronic word-of-mouth in hospitality and tourism management, Tourism Manag. 29 (2008) 458–468.
[3] E. Marchiori, L. Cantoni, The online reputation construct: does it matter for the tourism domain? A literature review on destinations' online reputation, Inform. Technol. Tourism 13 (3) (2011) 139–159.
[4] N. Kushcheva, Monitoring online reputation of tourist destinations in Finland, in: INTED2022 Proceedings, IATED, 2022, pp. 9442–9447.
[5] A. Arumugam, S. Nakkeeran, R. Subramaniam, Exploring the factors influencing heritage tourism development: a model development, in: 15, 2023, p. 11986.
[6] G. Dowling, Creating Corporate Reputations. Identity, Image, and Performance, Oxford University Press, Oxford, 2001.
[7] A. Inversini, E. Marchiori, C. Dedekind, and L. Cantoni, "Applying a conceptual framework to analyze online reputation of tourism destinations," 2010.
[8] Marco Cioppi, Ilaria Curina, Fabio Forlani, Tonino Pencarelli, Online presence, visibility and reputation: a systematic literature review in management studies, J. Res. Interac. Marketing (2019), https://doi.org/10.1108/JRIM-11-2018-0139 ahead-of-print.
[9] V. Cillo, R. Rialti, M.D. Giudice, A. Usai, Niche tourism destinations' online reputation management and competitiveness in big data era: evidence from three Italian cases, Curr. Issues Tourism 24 (2) (2021) 177–191, https://doi.org/10.1080/13683500.2019.1608918.
[10] L. Zhu, Y. Lin, M. Cheng, Sentiment and guest satisfaction with peer-to-peer accommodation: when are online ratings more trustworthy? Int. J. Hosp. Manag. 86 (2020), 102369.
[11] A. Crisci, V. Grasso, P. Nesi, G. Pantaleo, I. Paoli, I. Zaza, Predicting TV programme audience by using Twitter-based metrics, Multimedia Tools Applic. (2017) 1–30.
[12] D. Cenni, P. Nesi, G. Pantaleo, I. Zaza, Twitter vigilance: a multi-user platform for cross-domain Twitter data analytics, NLP and sentiment analysis, in: IEEE international Conference on Smart City and Innovation, San Francisco, 2017.
[13] A. Chauhan, K. Kummamuru, D. Toshniwal, Prediction of places of visit using tweets, Knowl. Inf. Syst 50 (1) (2017) 145–166.
[14] M. Hu, H. Li, H. Song, X. Li, R. Law, Tourism demand forecasting using tourist-generated online review data, Tourism Manag 90 (2022), 104490.
[15] G.E. Box, G.M. Jenkins, G.C. Reinsel, G.M. Ljung, Time Series analysis: Forecasting and Control, John Wiley & Sons, 2015.
[16] Zheng Xiang, Qianzhou Du, Yufeng Ma, Weiguo Fan, A comparative analysis of major online review platforms: implications for social media analytics in hospitality and tourism, Tourism Manag. (2017) 58, https://doi.org/10.1016/j.tourman.2016.10.001.
[17] M. Chu, Y. Chen, L. Yang, J. Wang, Language interpretation in travel guidance platform: text mining and sentiment analysis of TripAdvisor reviews, Front. Psychol. 13 (2022), https://doi.org/10.3389/fpsyg.2022.1029945.
[18] K. Puh, M. Bagić Babac, Predicting sentiment and rating of tourist reviews using machine learning, J. Hospital. Tourism Insights 6 (3) (2023) 1188–1204.
[19] T. Chen, C. Guestrin, XGBoost, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, https://doi.org/10.1145/2939672.2939785.
[20] C. Lea, R. Vidal, A. Reiter, G.D. Hager, Temporal convolutional networks: a unified approach to action segmentation, arXiv (2016), https://doi.org/10.48550/ARXIV.1608.08242.
[21] B. Lim, S.O. Arik, N. Loeff, T. Pfister, Temporal fusion transformers for interpretable multi-horizon time-series forecasting, arXiv (2019), https://doi.org/10.48550/ARXIV.1912.09363.
[22] S. Liu, Y. Zhao, F. Xue, B. Chen, X. Chen, DeepCount: crowd counting with WiFi via deep learning, arXiv preprint (2019) arXiv:1903.05316.
[23] H. Wu, C. Gao, Y. Cui, R. Wang, Multipoint infrared laser-based detection and tracking for people counting, Neural. Comput. Appl. 29 (2018) 1405–1416.
[24] E. Collini, L.A. Ipsaro Palesi, P. Nesi, G. Pantaleo, W. Zhao, Flexible thermal camera solution for smart city people detection and counting. Accepted For Publication On Multimedia Tools and Applications, Springer, 2023, https://doi.org/10.1007/s11042-023-16374-x.
[25] Z. Ivanovski, A. Milenkovski, Z. Narasanov, Time series forecasting using a moving average model for extrapolation of number of tourist, UTMS J. Economics 9 (2) (2018).
[26] Y.-W. Chang, C.-Y. Tsai, Apply deep learning neural network to forecast number of tourists, in: 2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA), Taipei, Taiwan, 2017, pp. 259–264, https://doi.org/10.1109/WAINA.2017.125.
[27] H. Laaroussi, F. Guerouate, M. sbihi, Deep Learning Framework for Forecasting Tourism Demand, in: 2020 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD), Marrakech, Morocco, 2020, pp. 1–4, https://doi.org/10.1109/ICTMOD49425.2020.9380612.
[28] Y. Chen, T. Hu, P. Song, Identifying the role of media discourse in tourism demand forecasting, Curr. Issues Tourism (2023) 1–15, https://doi.org/10.1080/13683500.2023.2165050.
[29] H. Li, M. Hu, G. Li, Forecasting tourism demand with multisource big data, Annals Tourism Res. 83 (2020), 102912, https://doi.org/10.1016/j.annals.2020.102912.
[30] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (4) (1943) 115–133.
[31] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learning 20 (3) (1995) 273–297.
[32] K. Cho, Bart Van Merri"enboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, arXiv Preprint arXiv (2014). 1406.1078.
[33] S. Hochreiter, J"urgen Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[34] Q.-T. Phan, Y.-K. Wu, Q.-D. Phan, A comparative analysis of XGBoost and temporal convolutional network models for wind power forecasting, in: 2020 International Symposium on Computer, Consumer and Control (IS3C), Taichung City, Taiwan, 2020, pp. 416–419, https://doi.org/10.1109/IS3C50286.2020.00113.
[35] X. Hu, Stock price prediction based on temporal fusion transformer, in: 2021 3rd International Conference on Machine Learning, Big data and Business Intelligence (MLBDBI), Taiyuan, China, 2021, pp. 60–66, https://doi.org/10.1109/MLBDBI54094.2021.00019.
[36] N. Chugh, N. Phumchusri, Bangkok tours and activities data analysis via user-generated content, in: 2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE), Southend, UK, 2020, pp. 98–102, https://doi.org/10.1109/iCCECE49321.2020.9231211.
[37] D. Asteriou, S.G. Hall, ARIMA models and the Box–Jenkins methodology, Appl. Econometrics 2 (2) (2011) 265–286.
[38] T.K. Ho, Random decision forests, in: Proceedings of 3rd international conference on document analysis and recognition 1, 1995, pp. 278–282.

[39] H. Song and L. Han, “Predicting tourist demand using big data,” 2017, pp. 13–29. doi:10.1007/978-3-319-44263-1_2.
[40] S.J. Miah, H.Q. Vu, J. Gammack, M. McGrath, A big data analytics method for tourist behaviour analysis, Inform. Manag 54 (6) (2017) 771–785, https://doi.org/10.1016/j.im.2016.11.011.
[41] M. De la Calle-Vaquero, M. García-Hernández, S. Mendoza de Miguel, Urban planning regulations for tourism in the context of overtourism. applications in historic centres, Sustainability 13 (1) (2021) 70.
[42] M.T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?”: explaining the predictions of any classifier, arXiv (2016), https://doi.org/10.48550/ARXIV.1602.04938.
[43] O. Tokarchuk, J.C. Barr, C. Cozzio, How much is too much? Estimating tourism carrying capacity in urban context using sentiment analysis, Tourism Manag. 91 (2022), 104522.
[44] S.F. McCool, D.W. Lime, Tourism carrying capacity: tempting fantasy or useful reality? J. Sustain. Tourism 9 (5) (2001) 372–388.
[45] A. Ogunleye, Q.G. Wang, XGBoost model for chronic kidney disease diagnosis, IEEE/ACM Trans. Comput. Biol. Bioinf. 17 (6) (2019) 2131–2140.
[46] F. Ramdani, M.T. Furqon, The simplicity of XGBoost algorithm versus the complexity of random forest, support vector machine, and neural networks algorithms in urban forest classification, F1000Research 11 (2022) 1069.
[47] M. Zou, W.G. Jiang, Q.H. Qin, Y.C. Liu, M.L. Li, Optimized XGBoost model with small dataset for predicting relative density of Ti-6Al-4V parts manufactured by selective laser *melting*, Materials 15 (15) (2022) 5298.
[48] J. Vainio, K. Holmberg, Highly tweeted science articles: who tweets them? An analysis of Twitter user profile descriptions, Scientometrics 112 (2017) 345–366.
[49] F. AmArAl, T. TiAgo, F. TiAgo, User-generated content: tourists’ profiles on Tripadvisor, Int. J. Strategic Innovative Market. 1 (3) (2014) 137–145.
[50] V. Arefieva, R. Egger, TourBERT: a pretrained language model for the tourism industry, arXiv preprint arXiv:2201.07449. (2022).
[51] Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding.
[52] C.W.J. Granger, Investigating causal relations by econometric models and cross-spectral, Methods Econom 37 (3) (1969) 424–438, https://doi.org/10.2307/1912791.
[53] G. De Luca, A. Shirvani Dastgerdi, C. Francini, G. Liberatore, Sustainable cultural heritage planning and management of overtourism in art cities: lessons from atlas world heritage, Sustainability 12 (9) (2020) 3929.