Social Media Data Analytics for Tourism
A Preliminary Study

Arkka Dhiratara, Jie Yang, Alessandro Bozzon, Geert-Jan Houben

Delft University of Technology, Mekelweg 4, 2628CD, Delft, The Netherlands
A.Dhiratara@student.tudelft.nl,
{J.Yang-3, A.Bozzon, G.J.P.M.Houben}@tudelft.nl

Abstract. Social media data are increasingly used as the source of research in a variety of domains. A typical example is urban analytics, which aims at solving urban problems by analyzing data from different sources including social media. The potential value of social media data in tourism studies, which is one of the key topics in urban research, however has been much less investigated. This paper seeks to understand the relationship between social media dynamics and the visiting patterns of visitors to touristic locations in real-world cases. By conducting a comparative study, we demonstrate how social media characterizes touristic locations differently from other data sources. Our study further shows that social media data can provide real-time insights of tourists' visiting patterns in big events, thus contributing to the understanding of social media data utility in tourism studies.

1 Introduction

Today, our society is increasingly hyper-connected with the unprecedented rise of social media; currently, social media has 2.206 Billion active users with 30% global penetration[1]. Social media activities are therefore an important class of daily activities performed by people worldwide to fulfill their social needs. These social media activities have generated a wealth of social data, which can provide meaningful and even possibly, real-time insights to a variety of studies. Social media has thus been leveraged as part of marketing strategy for industries, through "passive marketing" (as sources of market intelligence to gain insights of the users) [7]. As Mangold and Faulds (2009) recommend, social media should be regarded as an integral part of an organization's marketing strategy and should not be taken lightly [14].

Among different domains, urban science has been shown as an important domain where social media data can contribute [8, 5]. A wide range of urban problems have been studied with social media data, including event detection [13], urban area characterization [8], to mention a few. The potential of social media data, however, has been much less investigated in the study of tourism, despite the fact that tourism plays a key role in economic and social development

[1] http://wearesocial.com/uk/special-reports/global-statshot-august-2015

of many cities. Social media data are intrinsically different compared with other data sources used for tourism studies [18][11], such as visitor survey, transportation statistics, or online reviews. They either require a considerable amount of laborious effort for data acquirement – thus are infrequently updated, or require a large amount of volunteer input from online users, thus are highly sparse. Different from them, social media data are easily accessible in big size; moreover, they are topically, and spatially and temporally tagged, thus providing a distinct opportunity for tourism studies.

On the other hand, the high availability of social media data raises some challenges. In order to retrieve relevant social data, one should take into account effective parameters for data filtering, such as hashtags, keywords or geographic boundaries/coordinates. When it comes to tourism studies, the selection of parameters is crucial and needs to be carefully designed, to avoid biasing the interpretation of the results. For example, to capture the popularity of Eiffel Tower in Paris w.r.t. the number of visitors, we should avoid only using keywords or hashtags (e.g. #EiffelTower, #TourEiffel, or #Eiffel) to filter the data, as we cannot assume whether people posting tweets with these hashtags are indeed currently visiting that location. In this case, geographic coordinates become necessarily to be used as an additional parameter for filtering social media data.

By carefully selecting the parameters for filtering social media data, we created a dataset to explore the potential of social media data for tourism studies. Specifically, we focus on Instagram, which has been shown to be highly popular among tourists [22] as it features creating and sharing visual content (i.e. images) by users. We then seek to answer the following research questions:

RQ1. *How does social media data characterize touristic location differently from other data sources?*

RQ2. *Can social media data provide real-time insights about visiting patterns of tourists to different tourist locations in big events?*

We tackle these questions by quantitatively and qualitatively analyzing the social media data we collected from Instagram. By comparing the social data results with official tourism statistics and online review data, we find that social media characterizes touristic locations differently, featuring more landmark locations. Our study further shows that, during big events, social media data can characterize the visiting patterns of tourists at different locations with high temporal resolution, thus contributing to the understanding of how social media data can be used for obtaining real-time insights for touristic locations.

Our study provides preliminary however useful results that support social media as a useful data source for tourism studies. In 2014, World Travel Tourism Council (WTTC) stated that tourism industry has contributed US$7.6 trillion and also support 276 million jobs across the globe [21]. Because of the high economical and societal relevance, governments and the tourism industry are devoted to attracting more tourists to visit their cities' touristic locations [10]. The results of this work can be of fundamental interests for tourism marketing and decision making using social media data.

2 Related Works

Social media data analytics. The unprecedented popularity of social media has offered opportunities for a variety of domains. An important application is user modeling, for which social media data has been shown to be useful for modeling user attributes, including personal traits and personality [4], their interests [2], and etc.. In addition, social media data have also been shown to be effective for detecting social trends, including hot topics [15], cultural fashion trends [12] , and even epidemic burst [9]. Different from existing studies, we focus on tourism analytics using social media data in this work.

Social media data in urban analytics. While little work can be found in tourism studies, social data has enabled a wealth of research works in urban analytics. These include using social data for event-detection [13], venue recommendation for city-scale events [5], characterizing mobility patterns [6] in cities, and etc.. These works aim to show the effectiveness of social data for characterizing urban environment and human behaviour in cities. Falling in this line of research, the potential of social media in tourism studies, however, remains to be investigated.

Data sources for tourism studies. United Nations and European Union have provided guidelines and recommendation regarding the comprehensive methodological framework for collection and compiling of tourism statistics [18, 11]. Among different data sources for tourism study, visitor survey and transportation statistics are the ones used the most. In addition to this, some tourism studies focus on data from online review websites, e.g. TripAdvisor [17, 16, 3]. While being relevant, data from these review websites are highly sparse [3], due to the nature that they require volunteer input from online users. Recent studies [3] have shown a limited value of these data in tourism research.

3 Methods

In 2015, France was one of the top touristic destinations over the world [20], with more than 83 million tourist arrivals. In addition, the city of Paris itself also ranked as the most visited city in the world [19]. Therefore, this paper will scope the sample of touristic locations within the city of Paris (France). This section introduces our method in creating the dataset for answering our research questions, including the official tourism statistics, TripAdvisor statistics, and Instagram data.

3.1 Official Tourism Statistics and TripAdvisor

We are interested in understanding the characteristics of data from social media, compared with those from other data sources. To this end, we first determined the touristic locations considered in our study. We chose the top-5 touristic locations based on their popularity in Paris' annual visitor statistic [1] and in TripAdvisor. Table 1 reports the descriptive statistics of these locations, including

the annual visitors according to the official report, and the number of reviewers in TripAdvisor.

Comparing the popularity of touristic locations in the official report and in TripAdvisor, we can find that touristic locations have different popularity between the one reported by official statistics and the one observed through TripAdvisor. In particular, while there are more visitors in Cathedral than Eiffel Tower according to the official report, Eiffel Tower appears to be the one reviewed more often on TripAdvisor. In addition, different museums also show different popularity: Musée du Louvre is more popular than Musée d'Orsay in terms of visitors. However, people tend to review Musée d'Orsay more often on TripAdvisor. These observations clearly show a difference between the reality (official statistics) and statistics from the online review platform (TripAdvisor).

#	**Touristic Locations**	**Official Statistic**	**TripAdvisor**
		Visitors	Reviews
1	`Notre Dame Cathedral`	14,300,000	42,442
2	`Musée du Louvre`	9,134,000	58,648
3	`Eiffel Tower`	7,097,302	79,198
4	`Musée d'Orsay`	3,480,609	40,640
5	`Arc de Triomphe`	1,200,000	23,689

Table 1: Touristic locations in Paris: annual visitors reported by the official statistic, and number of reviewers in TripAdvisor.

3.2 Instagram Data

This work requires data acquisition from Instagram as the main object of study. Currently, Instagram has provided access to their API Endpoints[2] through which developers are able to interact with its data. Unfortunately, due to the change of Instagram's platform policy starting on 17 November 2015, every newly created app won't have full access of the API Endpoint (Sandbox Mode)[3]. Developers are required to submit their applications through a review process to be granted will full API access, which requires a long period of time. Therefore, we developed an alternative method here to track real-time posts available on Instagram, that was, using its search function.

Instagram's search function enables us to find photos based on keyword. Moreover, Instagram website provides "explore" feature that enables a user to explore trending and recent posts on a specified location based on the Location-

[2] `https://www.instagram.com/developer/`

[3] `http://developers.instagram.com/post/133424514006/instagram-platform-update`

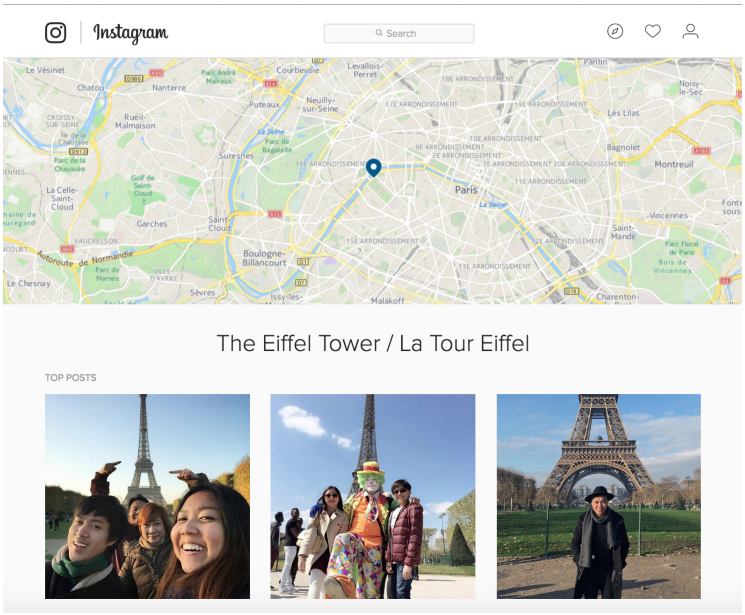


Fig. 1: Instagram's Explore page: The Eiffel Tower (Location-Id 218177821).

Id[4]. An example of the Location page is shown in Figure 1. By using web-scrapping techniques we were able to retrieve photos and auxiliary information required for this study, including Timestamp, User-Id and Location-Id.

Determine Relevant Location-Id. Instagram allows users to tag location information to photos. In doing this, users get automatically location tag recommendations based on their locations or photos' coordinates. We leveraged such location tags to filter relevant posts for this study. Particularly, we noticed that a single location can have multiple Location-Ids. For example, if we want to find Instagram's posts located at The Eiffel Tower we could find it on Location-Id 216052603, as well as Location-Id 2593354, which is the French version of the name Tour Eiffel. To cope with this problem, we proposed to merge multiple Location-Ids for individual touristic locations, in order to gain more data and to avoid biasing the interpretation of the results: including either only the English and the French version of the Location-Id of The Eiffel Tower, the results will be biased to specific type of people speaking English or French.

The discovery of available Location-Ids was performed as follows. First, we queried Instagram's posts based on keyword and/or hash-tags. Second, based on the retrieved posts we were able to identify a list of frequently used Location-Ids

[4] E.g., The Eiffel Tower: `https://www.instagram.com/explore/locations/218177821/`.

(> 100 posts) for specific touristic locations. As a result, we were able to find relevant Instagram's Location-Ids for each tourist destination considered in our study.

Tapping on Instagram's Recent Media Stream. After identifying the Location-Ids of every touristic location in our study, we built a crawler to scrap Instagram's explore page[5] using Location-Ids as the parameters. As a result, we were able to retrieve 898,339 Instagram posts. However, when analyzing the retrieved data based on post's timestamp, we found that the number of photos older than one month decreases dramatically. This pattern could happen either because of Instagram's own policy that not revealing all old photos or because of the feature itself only just available since June 2015[6].

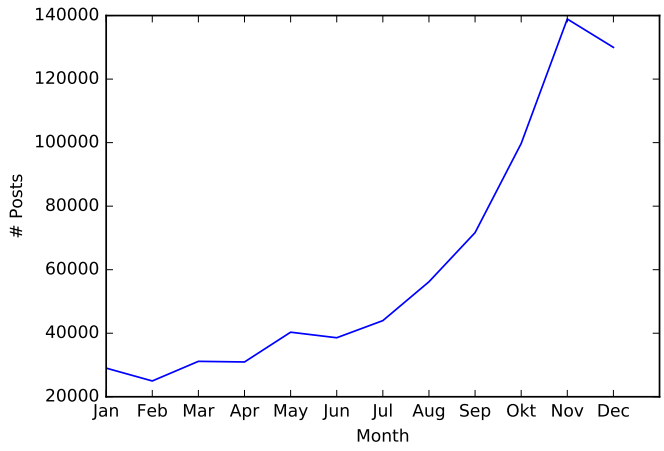


Fig. 2: Crawled Instagram's posts per month in 2015.

As a result of this finding, we used only recent retrieved posts that were not older than 14 days (from 25 December 2015 to 7 January 2016) to ensure reliability. We leave it to future work, to scrap the website for a longer period of time. The resulting data set contains 82,381 posts[7].

[5] `https://www.instagram.com/explore/locations/<location-id>`
[6] `http://blog.instagram.com/post/122260662827/150623-search-and-explore`
[7] The dataset has been made available to the community at: `https://github.com/arkka/tourism-analytics`

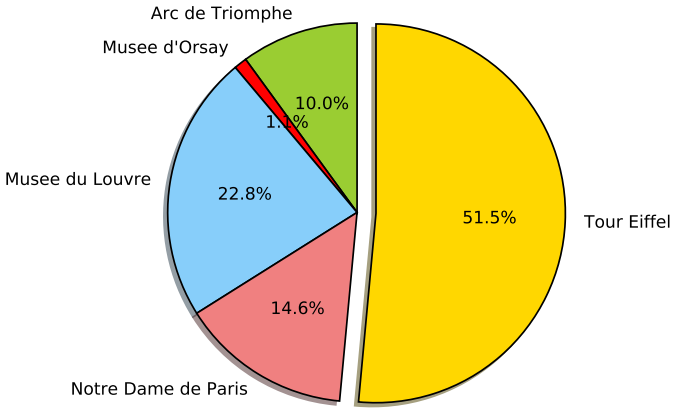


Fig. 3: Top tourist locations in Paris based on Instagram's posts.

4 Results

We now discuss our findings corresponding to the research questions.

4.1 How does Social Media Data Characterize Touristic Locations Differently from Other Data Sources?

In order to answer **RQ1**, we first construct a ranking list of the considered touristic locations according to their popularity in Instagram. Specifically for each touristic location, we count the number of related posts, to be used as a measure of its popularity in social media. Next, we compare this ranking list with the one in official tourism report and that in TripAdvisor.

Comparing Metrics. To quantitatively compare the ranking list of touristic locations from different data sources, we used a well-accepted measure of non-parametric rank correlations, namely Kendall's τ rank correlation, which is based on pair-wise agreements between touristic locations. The result shows that there is no correlation (< 0.1, with p-value > 0.1) between the ranking list of official report and that of social media. Comparing with the ranking from TripAdvisor, we find a moderate correlation between social media and TripAdvisor ($= 0.59$, p-value $= 0.14$). We next qualitatively analyze the similarity and dissimilarity among data from different sources, as we will see later.

Social Media vs. Official Statistic. It is interesting to see how the relative ranking positions of different tourist locations differ in social media list and the official list of tourism statistics. In particular, official tourism statistics reveals

that Notre Dame Cathedral and Muse du Louvre have more visitors than The Eiffel Tower, however in social media the most popular one is The Eiffel Tower for both Instagram and Tripadvisor. A possible reason for this would be that tourists are more opt to take a photo on landmark touristic locations without entering the attraction such as museums or monuments. Following this kind of scenario, we are able to leverage social media data to complement existing official statistics. Social media is able to identify trends independently without the needs of conventional visitor counting methods, such as using ticket sales or gate counters. Moreover, social media also able to provide high temporal resolution data that we can leverage to enhance our analysis which we explore in the later section.

Instagram vs. TripAdvisor. Comparing the popularity of touristic locations between TripAdvisor's user reviews and Instagram's posts, we find that there are certain amount of similarities between these statistic . Both of these statistics rank The Eiffel Tower as the most popular touristic locations, followed by Musée du Louvre and Notre Dame Cathedral respectively. The difference can be found for Musée d'Orsay and Arc de Triomphe, which respectively rank at the fourth and fifth in TripAdvisor, but it is the other way around on Instagram. Overall, it is interesting knowing the fact that these two different data sources can relate with each other. The similarity between TripAdvisor and Instagram data could be because that both are user-generated. In the following, we will analyze in detail the utility of social media data in providing fine-grained temporal insights for tourism studies.

4.2 Evolution of Touristic Location Popularity During New Year's Eve

Social media has been shown to be effective as sensors for detecting collective trends of societal events [7, 9]. However, the potential for real-time event sensing is less investigated. We now study **RQ2**, that is, how social media can reflect real-time dynamics of the popularity of touristic locations in big events. Specifically, we focus on the social media activities around different touristic locations in New Year's Eve and investigate how their popularity change over time.

Figure 4 shows the dynamics of the popularity of the top touristic locations in Paris within two weeks from Dec. 25th, 2015 to Jan. 7th, 2016 on a daily base. We can observe different visiting patterns of these considered touristic locations, according to the evolution of popularity before, during, and after New Year's Eve. In particular, the popularity of The Eiffel Tower and Arc de Triomphe drastically increases during New Year's Eve in terms of the number of Instagram posts. In contrast, the popularity of Musée du Louvre deceases during New Year's Eve. Therefore with social data we could draw the conclusion that The Eiffel Tower and Arc de Triomphe are the ones attract more social activities from visitors during the big event, while Musée du Louvre loses social attention during the same period of time. A possible reason could be that people are more likely

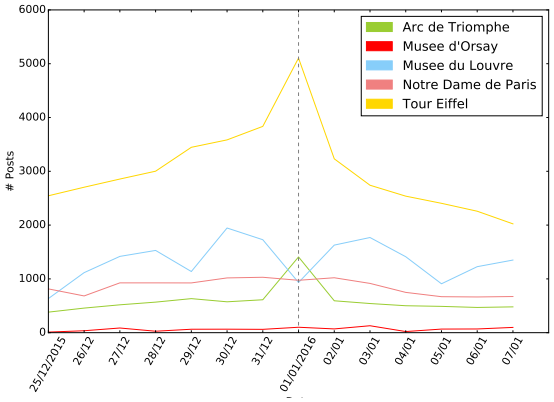
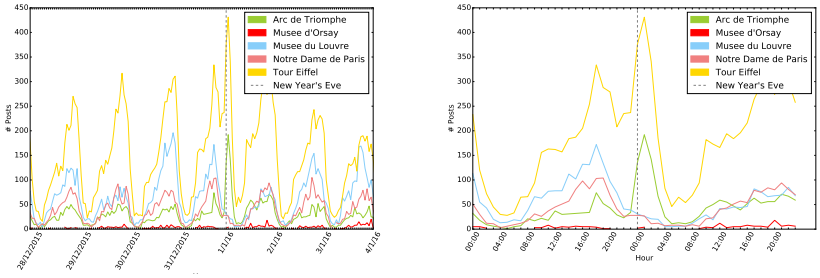


Fig. 4: Daily number of visitor trends on 2016 New Years Eve.



(a) Dec. 25th, 2015 – Jan. 7th, 2016 (b) 2016 New Year's Eve.

Fig. 5: The popularity of touristic locations over time, aggregated in each hour.

to celebrate big events in landmark attractions (e.g. The Eiffel Tower and Arc de Triomphe) than museums. We leave it to the future work to find out more characteristics of touristic locations according to their corresponding visiting patterns during big events.

4.3 Daily Visiting Patterns of Tourist Locations during New Year's Day

Social media data allow us to analyze fine-grained patterns of user activities with high temporal resolution. We now analyze the visiting patterns of people to touristic locations on hourly base.

Daily Visiting Pattern. Figure 5a shows the visiting frequency of people to touristic locations during the week centred at New Year's Day, observed through

Instagram. we could observe that in normal days (excluding New Year's Eve) number of Instagram's posts peaked at 9-10PM on each day. In contrast, we have found a drastic change of daily visiting pattern on New Year's Eve, which is moved earlier to 6PM on 31 December 2015 and followed by another peak on the next day at 1AM with a huge number of posts compared to the rest of other peaks. After the New Year's Eve (1 January 2016) we could observe that the daily visiting pattern comes back normal as usual in the evening, which is peaked at 9PM. These observations indicate a distinct visiting pattern of people to touristic locations on New Year's Day, differing from other days before and after. Moreover, this figure also shows the distribution of each post by touristic locations. Overall, The Eiffel Tower consistently has the highest number of posts compared others, which followed by Musée du Louvre and Notre Dame Cathedral respectively.

Using high temporal data that we have retrieved from social media, we are able to relate the number of posts on specific touristic locations with their opening hours. For example, the opening time of Musée du Louvre is from 9am to 6pm every day except on January 1, May 1 and December 25[8]. Based on the hourly visitor trends depicted on figure 5b, we could see that the number of posts increases dramatically on 9am until reaching it's peak on 6pm and then decreases dramatically when reaching the closing hour on 6pm. We could interpret that as during New Year's eve, people tend to visit a touristic location that has no restrictive opening hours, which usually is an outdoor touristic location, such as The Eiffel Tower and Arc de Triomphe, instead of museums in Paris such as Musée du Louvre that are closed on the 1st January 2016. Therefore, it is clear that social data is able to provide additional data and insights that complement the existing official statistics.

Visiting Patterns of Different Touristic Locations on New Year's Eve. Based on the previous analysis, we have noticed that there is a difference visiting pattern on New Year's Eve and normal days. On figure 5b, we focused our analysis from 31 December 2015 to 1 January 2016 to identify the trend. Consistent with previous results, this figure shows that on New Year's Eve people are more likely to celebrate at The Eiffel Tower and Arc de Triomphe. Moreover, Arc de Triomphe become the second-most popular touristic location for celebrating New Year's Eve. In contrast, Musée du Louvre which is usually the second-most popular touristic location on a normal day becomes less popular on New Year's Eve. These observations indicate that people tend to be selective in celebrating the big event, favouring more at landmark locations than museums.

5 Threat to Validity

This study aims to explore the potential of social media data in tourism studies. The main limitation (and a threat to validity) is the size of the data we collect.

[8] http://www.louvre.fr/en/hours-admission/admission

This relates to several important factors, including the length of the time period of the data, the number of city and touristic locations we consider in the data. It would be highly interesting to compare the touristic locations in different cities, which we leave for the future work. Second, due to Instagram's Public API limitation, we couldn't explore the full potential of Instagram data using the API. In this paper, we stress our effort in creating the trustable data crawler, and testing the reliability of the data.

6 Conclusions

While social media has open many research opportunities in urban analytic, the value in the tourism studies has been much less explored and remains an open question. This paper takes a first step towards understanding the utility of social media data in urban tourism studies. We show that social media can provide different reflections of the tourism from other data sources, including official tourism report and online review platforms (e.g. TripAdvisor); moreover, we find that social media data can provide real-time insights of tourists' visiting patterns during big events. In conclusion, our study provides a set of preliminary results that support social media as a useful data source for touristic marketing and decision making.

References

1. *Frequentation Des Sites Culturels Pariens 2015*. 2016.
2. F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *Extended Semantic Web Conference*, pages 375–389. Springer, 2011.
3. J. K. Ayeh, N. Au, and R. Law. do we believe in tripadvisor? examining credibility perceptions and online travelers attitude toward using user-generated content. *Journal of Travel Research*, page 0047287512475217, 2013.
4. Y. Bachrach. Personality and patterns of facebook usage. 2012.
5. M. Balduini, A. Bozzon, E. Della Valle, Y. Huang, and G.-J. Houben. Recommending venues using continuous predictive social media analytics. *IEEE Internet Computing*, 18(5):28–35, 2014.
6. Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. In *Proceedings of the 5th AAAI conference on web and social media*, volume 2011, pages 81–88, 2011.
7. E. Constantinides. Social media / web 2.0 as marketing parameter: An introduction. 2009.
8. J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the 6th AAAI Conference on Weblogs and Social Media*, page 58, 2012.
9. E. Diaz-Avilez and A. Stewart. Tracking twitter for epidemic intelligence. case study: Ehec/hus outbreak in germany. 2011.
10. M. S. et al. Tourist satisfaction as the key to destination survival in pahang. 2013.
11. Eurostat. *Methodological manual for tourism statistics*, 2014.

12. G. C. J. Park and E. Ferrara. Sytle in the age of instagram: Predicting success within the fashion industry using social media. 2015.
13. R. Lee and K. Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*, pages 1–10. ACM, 2010.
14. W. Mangold and D. Faulds. Social media: The new hybrid element of the promotion mix. 2009.
15. M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1155–1158. ACM, 2010.
16. J. Miguéns, R. Baggio, and C. Costa. Social media and tourism destinations: Tripadvisor case study. *Advances in Tourism Research*, 26(28):26–28, 2008.
17. P. O'Connor. User-generated content and travel: A case study on tripadvisor. com. *Information and communication technologies in tourism 2008*, pages 47–58, 2008.
18. United Nations. *International Recommendations for Tourism Statistics*, 2008.
19. UNWTO. *UNWTO Tourism Highlights: 2015 Edition*. 2015.
20. UNWTO. Unwto world tourism barometer. 14, May 2016.
21. WTTC. *Travel & Tourism: Economic Impact 2015*. 2015.
22. J. Yang, C. Hauff, G.-J. Houben, and C. T. Bolivar. Diversity in urban social media analytics. In *Web Engineering*, pages 335–353. Springer, 2016.