# Analysis and Knowledge Extraction from Event-related Visual Content on Instagram

Tahereh Arabghalizi, Behnam Rahdari, and Marco Brambilla

Politecnico di Milano, Via Ponzio, 34/5, 20133 Milano, Italy {tahereh.arabghalizi,behnam.rahdari}@mail.polimi.it, marco.brambilla@polimi.it

**Abstract.** Nowadays people share everything on online social networks, from daily life stories to the latest local and global news and events. Many researchers have exploited this as a source for understanding the user behaviour and profile in various settings. In this paper, we propose two quantitative methods that investigate the relevance of the published photos about a cultural event in terms of knowledge that can be extracted, user behaviour and relation to the context of the event. We show our approach at work for the monitoring of participation to a large-scale artistic installation that collected more than 1.5 million visitors in just two weeks (namely *The Floating Piers*, by *Christo and Jeanne-Claude*). We report our findings and discuss the pros and cons of the analysis.

Keywords: Social Media, Big Data, Image Analysis

## 1 Introduction

Today social networks are the most popular communication channels for users looking to share their experiences and interests. They host considerable amounts of user-generated materials for a wide variety of real-world events of different type and scale [5]. Social media has a significant impact in our daily lives. People share their opinions, stories, news, and broadcast events using social media. Monitoring and analyzing this rich and continuous flow of user-generated content can provide valuable information, enabling individuals and organizations to acquire insightful knowledge [6]. Due to the immediacy and rapidity of social media, news events are often reported and spread on Twitter, Instagram, or Facebook ahead of traditional news media [8].

Despite the importance of social media, the number of studies and analyses on the impact of cultural and art events in social networks is rather limited, and focused on English-only content or are tailored to only one specific site, with addressing one type of document e.g., textual messages, photos or videos. Moreover, due to the noisy nature of the data extracted from social media, especially ungrammatical and ambiguous textual features, previous works [1, 11] proposed a comprehensive preprocessing method that normalizes and translates

#### 2 Tahereh Arabghalizi, Behnam Rahdari, and Marco Brambilla

texts to make the data clean and consistent. However, this technique might not be useful in Instagram which is known as a photo-sharing platform.

In this paper we aim to analyze visual social media content specifically photos related to a cultural or art event on Instagram. We capture the visual features of photos (namely colors, concepts, and demographics of people), we extract contextual and behaviour knowledge about what and how users share about the event, and then based on this we can tackle our main research questions: (1) finding the relevance between the shared photos about an event and the event itself, and (2) extract a summary of the statistics of the event and its attendees. Our findings can help marketing and event organizers in creating engaging content that communicates more effectively with their audiences and their future customers.

The paper is organized as follows: Section 2 discusses the related work; Section 3 describes our methods and data; Section 4 reports the outcomes of the analysis. Finally, Section 5 concludes and outlines the future work.

# 2 Related Work

Several recent researches proposed techniques for identifying social media content for planned events. Many of these approaches like [13] are limited in the amount and types of event content that they can handle. In other words, they rely on known event content in the form of manually selected terms from a single social media site, while a most related research [4] focuses on identifying meaningful event-related concepts, across multiple social media sites namely Twitter, YouTube, and Flickr, with varying types of documents (e.g., texts, videos, photos). Becker at el. [4] presented a query-oriented solution to automatically retrieve social media documents for any known event, without any assumption about the textual content of the event or its associated documents.

In recent years, creating effective content for social media marketing campaigns has become a challenge to understand what drives user engagement. While researchers have applied various methods to study how users engage with textual [10, 12], only a few have also focused on and visual content [14, 9]. Jaakonmäki at el. [9] reports on a quantitative study that extracts textual and visual content features from Instagram posts to statistically model their influence on user engagement. Among the work that address the visual content in social media, some aim to infer users' personality traits and viewers' engagement from the shared photos and their applied filters [7, 3, 2]. For instance, Bakhshi at el. [2] studied the engagement value of photos with human faces in them. They found that photos with faces are more likely to receive likes and comments.

In contrast with these efforts, we focus on analyzing the different aspects of event-related visual content on Instagram and show it at work on a real case study.

### 3 Methods and Data

Our main objective in this work is to exploit the knowledge that can be extracted via low-level and high-level features of shared images for finding the relevance between the shared photos about an event and the event itself. We follow two quantitative approaches to investigate the relationship between content features of Instagram photos and a cultural or art event.

The first approach employs the concepts (i.e., objects or entities detected in the image) that can be extracted from photos to find the level of relevance of the image; based on this, we classify the images into two classes, as relevant and irrelevant.

The second method finds relevant images by analyzing the color schema of each photo and specifying the relevance based on existence of the main color pattern(s) related to the event.

In this section, we describe how we collected and analyzed the data, and present a statistical overview of our case study.

### 3.1 Case Study and Data Extraction

This study exploits Instagram and Twitter datasets from a famous artwork called "The Floating Piers" that was created by the world-renowned artists Christo and Jeanne-Claude <sup>1</sup> and exposed to the public view at the Lake Iseo in Italy, from June 18 through July 3 2016 (see Figure 1<sup>2</sup>).



Fig. 1. The Floating Piers by Christo and Jeanne-Claude

We use this artistic event as a use case for our methods. We extracted the social media content relevant to the event, during a time period from June 10th

<sup>&</sup>lt;sup>1</sup> http://christojeanneclaude.net/projects/the-floating-piers

<sup>&</sup>lt;sup>2</sup> Photo Credits:Sailko, Monte Isola. License: Creative Commons Attribution-Share Alike 3.0 Unported.

#### 4 Tahereh Arabghalizi, Behnam Rahdari, and Marco Brambilla

to July 30th 2016, that contains 30,256 Instagram posts and 14,062 tweets, using Twitter and Instagram APIs.

Figure 2 illustrates the total numbers of Instagram posts vs. tweets within a timeline. One could conclude that Twitter users have a tendency to tweet about the news at the moment when an event starts, whereas Instagram users usually share their experiences when an event ends.

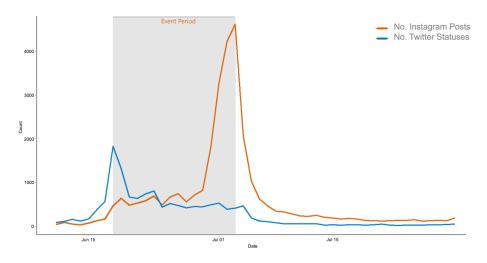


Fig. 2. Time series of Tweets vs. Instagram posts

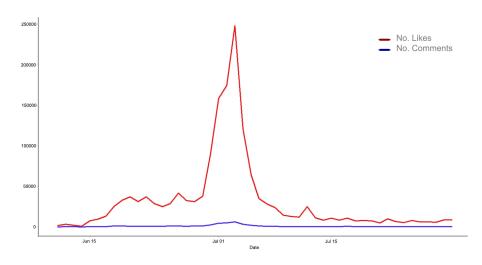
#### 3.2 Overview of the Event in Instagram

To have a clear intuition of the level of user engagement in Instagram, the volume of likes and comments received by uploaded posts are depicted in Figure 3. As demonstrated, Instagram users are more interested in liking the posts rather than commenting, that is why the number of comments is much less than likes count and remains on a constant rate during the time interval.

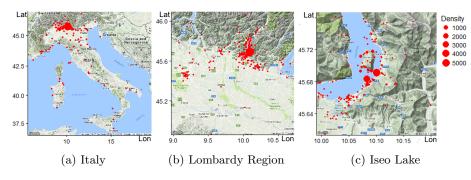
According to the statistics, unlike Instagram users, most Twitter users are not willing to specify the location of their published tweets. We displayed the density of Instagram posts on geographical plots in Figure 4. As one can see the density of posts has a direct relationship with their locality which means most Instagram posts have been published near the main venue of the event.

#### 3.3 Quantitative Methods

Our research process continued with collecting a random sample of Instagram posts (3000) because of the limitation of requests in Clarifai API. Then we captured and stored available visual features namely concepts, colors schema



 ${\bf Fig.~3.}$  Instagram total likes vs. comments



 ${\bf Fig.~4.}$  Density of Instagram posts in different coordinates

6 Tahereh Arabghalizi, Behnam Rahdari, and Marco Brambilla

and demographic features of people (faces) in photos including age, gender and race, using Clarifai.

In order to evaluate our proposed methods, we designed a web-based survey<sup>3</sup> consisting of two questions about each Instagram photo: 1- Is this photo related to the Floating Piers event? 2- Does this photo contain the Piers? We asked three people to answer these questions for all 3000 photos that we had in the dataset.

In the first approach, we try to find the relationship between the event and the concepts in the photos that are captured by Clarifai. Theoretically speaking, if the concepts found in the photos are similar to the real concepts of the event, we can conclude that those photos are related to the event and thus are not spams. To make this method quantitative, we assign a numerical weight to each concept which is its normalized frequency (number of repetitions) in the set of photos. This way the most frequent concepts (e.g., travel, water, sea, outdoors) gain higher weights than other words. Subsequently, we sum all the weights corresponding to a photo to calculate the final score of that photo. After finding the right threshold for this score, we determine which photos belong to the event. In the end, we compare the results of the survey and this method by computing performance measures that will be explained in section 4.2.

In the second approach, we try to find the relationship between the event, in particular the piers' structure, and the top colors in the shared photos that can be extracted by Clarifai. To recognize the presence of the Floating Piers artifacts in the photos, we search through all extracted colors of each photo and check if there are any colors in a specific shade (the piers' color shade). Then we compare the results of the survey and this method by computing performance measures that will be explained in section 4.2.

# 4 Results and Discussion

In this section, the most significant results of the experiment over the case study are shown and discussed.

#### 4.1 Dataset-related Results

Using Clarifai API, we can exract the number of faces (people) in each photo and each person's dempgraphic features such as gender, age and race. As presented in Figure 5, nearly 75 percent of shared photos do not include a face (person) while 12 and 14 percent of photos include one person and a group (two or more persons) respectively. However, the avergae number of likes and comments that photos containing one person gained is almost equal to the avergae number of likes and comments of the majority of photos (with no face). One can conclude that portraits (and selfies) receive more attention from users in Instagram.

According to the data extracted from Clarifai, approximately both female and male equally participated in the event (50.4%, 49.6%). Moreover, as shown

<sup>&</sup>lt;sup>3</sup> https://goo.gl/etvZqM

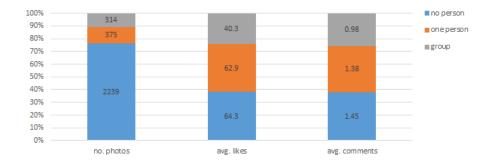


Fig. 5. Average number of likes and comments for photos with no person, one person and a group

in Figure 6 and Figure 7, three quarter of attendees were between 25 and 45 years old and 67 percent of them were white.

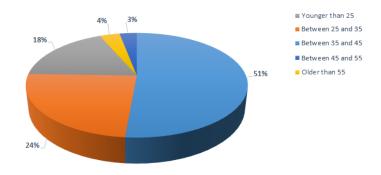


Fig. 6. Age distribution of the event attendees

One of the most popular features of Instagram is that it allows its users to capture and customize their photos and videos with several filter effects. Considering that, we extracted the filters applied on photos to see if the users were interested in using filters for their photos taken from The Floating Piers or not. The results are indicated in Figure 8 and shows that more than half of the photos were uploaded on Instagram with no filter.

### 4.2 Approach-related Results

As explained in section 3.3, in the first method we extracted the concepts of each photo using Clarifai API and then computed the relevance scores. Figure 9.a shows the most frequent concepts (words) appeared in all photos. Besides,

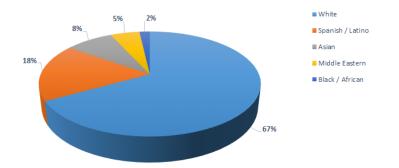


Fig. 7. Race distribution of the event attendees

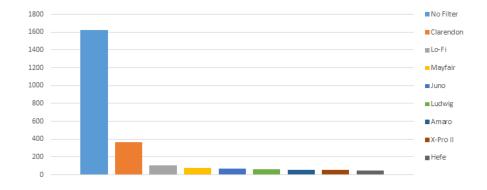
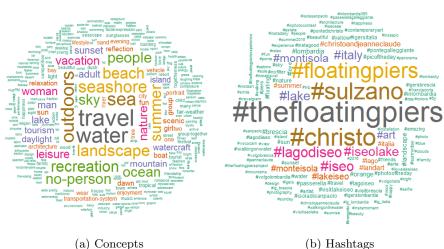


Fig. 8. Top filters vs. the number of photos  $% \mathcal{F}(\mathcal{F})$ 

in favour of comparison between these concepts and user generated content, we extracted the hashtags of each photo using Instagram API (Figure 9.b). As it can be seen Instagram users, in this event, do not usually tend to use hashtags to describe their shared photos using existing concepts in the photos.



(a) Concepts

Fig. 9. Word cloud representations

Subsequently, in order to find the right threshold for the calculated relevance scores, we use *discrete derivative* which is an analogue of derivative for a function (here the descending order of scores) whose domain is discrete. As can be seen in Figure 10, the value of the discrete derivative is maximum when the relevance score is 2.4. So we set the threshold to this number and consider all the photos with scores lower that this threshold as irrelevant.

As mentioned earlier, in the second method we extracted top colors of each photo and then we used a specific color shades to distinguish between photos comprising the piers and the rest. As shown in Figure 11, the shades of orange are the biggest portions among the four main ranges of the colors, which makes sense because the color of the fabric used to make the piers is also in this color spectrum.

Once we have built our methods or models, the most important question that arises is how good they are. Therefore, to evaluate our methods we use *Confusion* Matrix in which true condition corresponds to the survey results and predicted condition corresponds to the outcomes of our proposed methods. Considering this matrix that is often used to describe the performance of a classification model, we calculate precision, recall and accuracy measures for each method separately and indicate their values in Table 1.

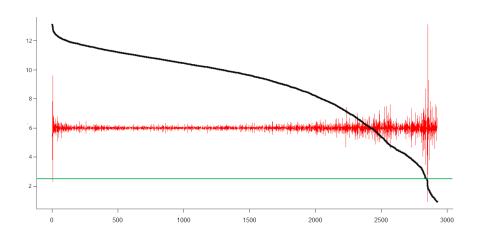


Fig. 10. Finding threshold for the relevance scores

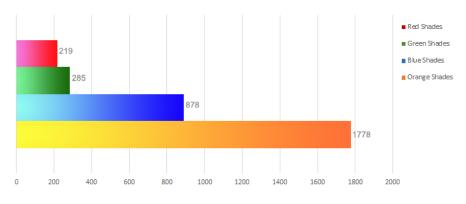


Fig. 11. Main color shades among all photos

Metric   Method 1 (Concepts) Method 2 (Colors)		
Precision Recall Accuracy	0.958 0.956 0.924	$\begin{array}{c c} 0.923 \\ 0.919 \\ 0.863 \end{array}$

 Table 1. Precision, recall and accuracy for two proposed methods

As one can see in this table, the accuracy of the first method is higher than the second one. Since our datasets are symmetric, which means that the values of false positive and false negative are almost the same, we can conclude that model with higher accuracy is a better model in terms of performance. Besides, the higher values of precision and recall for the first method are approved seals on the preference of this method.

### 5 Conclusion and Future Work

In this study, we proposed two quantitative methods to probe the relationship between features of Instagram photos and a cultural or art event and then employed an online survey to evaluate these methods. We used The Floating Piers event as a case study to show how the proposed approachs work with the real life scenarios.

Based on the outcomes of these two approaches we can conclude that employing concepts of photos (first method) eventuates more accurate results rather than using the extracted colors (second method). The reason behind that can be the high diversity of images in terms of angle of photography, time of the day, usage of Instagram filters etc., which can led to less precise analysis over colors. Furthermore, the resemblance of piers' color and other objects namely faces, foods, etc. in a picture can be another reason for the lack of accuracy in the second approach.

The current study can go further with considering other social media platforms such as Facebook, Google+, Flickr, etc. that might result in a clearer and wider picture of the characteristics of the event.

# References

- 1. ARABGHALIZI, T., RAHDARI, B.: Event-based user profiling in social media using data mining approaches (2017)
- Bakhshi, S., Shamma, D.A., Gilbert, E.: Faces engage us: Photos with faces attract more likes and comments on instagram. In: Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems. pp. 965–974. CHI '14, ACM, New York, NY, USA (2014), http://doi.acm.org/10.1145/2556288.2557403
- 3. Bakhshi, S., Shamma, D.A., Kennedy, L., Gilbert, E.: Why we filter our photos and how it impacts engagement. In: ICWSM. pp. 12–21 (2015)
- Becker, H., Iter, D., Naaman, M., Gravano, L.: Identifying content for planned events across social media sites. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. pp. 533–542. WSDM '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2124295.2124360
- Becker, H., Naaman, M., Gravano, L.: Learning similarity metrics for event identification in social media. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. pp. 291–300. WSDM '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1718487.1718524
- Farzindar, A., Wael, K.: A survey of techniques for event detection in twitter. Comput. Intell. 31(1), 132-164 (Feb 2015), http://dx.doi.org/10.1111/coin. 12017

- 12 Tahereh Arabghalizi, Behnam Rahdari, and Marco Brambilla
- Ferwerda, B., Schedl, M., Tkalcic, M.: Predicting personality traits with instagram pictures. In: Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015. pp. 7–10. EMPIRE '15, ACM, New York, NY, USA (2015), http://doi.acm.org/10.1145/2809643.2809644
- 8. Hu, Y.: Event Analytics on Social Media: Challenges and Solutions. Ph.D. thesis, Arizona State University (2014)
- Jaakonmäki, R., Müller, O., Brocke, J.v.: The impact of content, context, and creator on user engagement in social media marketing. In: 50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017 (2017), http://aisel.aisnet.org/hicss-50/da/data\_ text\_web\_mining/6
- Jamali, S., Rangwala, H.: Digging digg: Comment mining, popularity prediction, and social network analysis. In: Proceedings of the 2009 International Conference on Web Information Systems and Mining. pp. 32–38. WISM '09, IEEE Computer Society, Washington, DC, USA (2009), http://dx.doi.org/10.1109/WISM.2009. 15
- Rahdari, B., Arabghalizi, T., Brambilla, M.: Analysis of online user behaviour for art and culture events. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. pp. 219–236. Springer, Cham (2017)
- Sabate, F., Berbegal-Mirabent, J., Cañabate, A., Lebherz, P.R.: Factors influencing popularity of branded content in facebook fan pages. European Management Journal 32(6), 1001–1011 (2014)
- Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: Real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web. pp. 851–860. WWW '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1772690.1772777
- Yuheng, H., Lydia, M., Subbarao, K.: What we instagram: A first analysis of instagram photo content and user types, pp. 595–598. The AAAI Press (2014)