Analysis of Online User Behaviour for Art and Culture Events

Behnam Rahdari, Tahereh Arabghalizi, Marco Brambilla

Analysis of Online User Behaviour for Art and Culture Events

Behnam Rahdari, Tahereh Arabghalizi, and Marco Brambilla

Politecnico di Milano,
Via Ponzio, 34/5, 20133 Milano, Italy
{behnam.rahdari,tahereh.arabghalizi}@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 address the specific problem of user behavioural profiling in the context of cultural and artistic events. We propose a specific analysis pipeline that aims at examining the profile of online users, based on the textual content they published online. The pipeline covers the following aspects: data extraction and enrichment, topic modeling, user clustering, and prediction of interest. 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 work.

Keywords: Social Media, Big Data, Behaviour Analysis, Data Mining

1 Introduction

1.1 Context

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 [2]. 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 [8].

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 [15]. With the fast growth of social media, Twitter has become one of the most popular platforms for people to post short messages. Events like breaking news can easily draw peoples attention and spread rapidly on Twitter.

Therefore, the popularity and significance of an event can be measured by the volume of tweets covering the event. Furthermore, the relevant tweets are also indicators of opinions and reactions to events [6].

Obtaining demographic information about social media users, their interests and their behaviour is the main concern of *user profiling*, which in turn can be used to understand more about users and improve their satisfaction [16].

Various research works that have been conducted in user profiling, for instance in the field of recommender systems. However, the number of studies and analyses on the impact of cultural and art events in social media is rather limited, and focused on English-only content,while overlooking the other languages. Considering this, we propose a **domain-specific approach to profile social media users engaged in a cultural or art event**, regardless of their language and their location.

1.2 Problem Statement

In this study, we intend to respond the following questions:

- What are the **topics of interest of the social media users** who published their experiences or opinions about a cultural or artistic event?
- What **demographic features** can be revealed about these users?
- What is the **predicted level of engagement and areas of interest** of perspective users approaching the event?

To tackle the above questions, we suggest an approach that addresses two core aspects:

1. **User profiling:** the process of extracting user features, raising the level of abstraction of the discussed concepts, and deriving the topics of interest. the interest domains and behaviour of social media users who share their opinions about a cultural or artistic event.
2. **User interest prediction:** the anticipation of whether a social media user will be attracted by the current or similar event in the (near) future and, if yes, with what kind of interest and background.

1.3 Proposed Solution

The first step of our approach is to collect the required data about an event from social media. After cleaning and transforming this collected data to a proper format, we define some steps to perform data analysis in different levels. The first step of analysis is to extract the main topics from the provided dataset by using topic modeling techniques. After that, we perform different clustering algorithms on the outputs of topic modeling and then employ cluster validation techniques to evaluate the obtained results. Ultimately, using the outcomes of data analysis, we can employ a classification method to anticipate the interest areas of the future users in similar events.

Notice therefore that in our specific setting we cluster users by topics of interest, and not merely based on lexical similarity based on used words.

1.4 Structure of the paper

The paper is organized as follows: Section 2 discusses the related work; Section 3 describes our approach, with practical implementation details reported in Section 4; Section 5 presents the case study and Section 6 reports the outcomes of the analysis. Finally, Section 7 concludes and outlines the future work.

2 Related Work

Knowledge Discovery in Databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data. The KDD process involves using the database along with any required selection, preprocessing, subsampling, and transformations; applying data-mining methods (algorithms) to enumerate patterns from it; and evaluating the products of data mining to identify the subset of the enumerated patterns that are deemed useful for increasing knowledge [9].

Past works have found that content extracted from social media is a meaningful reflection of the human behind the social network account posting that content. Works like [26] and [28] mainly focus on clustering web users, while studies such as [22, 10, 1] specifically address clustering of people in social networks based on textual and non-textual features. There are also several works that address user profiling in online social networks. For instance, in [3] the authors propose a method to select experts within the population of social networks, according to the information about their social activities.

Other works [21, 18] focus on event analysis in social media. The former analyzes the resulted heterogeneous network, and use it in order to cluster posts by different topics and events; while the latter performs analysis and comparison of temporal events, rankings of sightseeing places in a city, and study mobility of people using geo-tagged photos. Some works [11] leverage Twitter lists to extract the topics that users talk about and [25] introduces validation methods to evaluate clustering results. All these works have delivered new solutions to social media analysis field and investigated the problems of profiling users and analyzing events by employing different data mining approaches.

In comparison to the mentioned studies, our research proposes a complex approach that aims clustering social media users based on their topic of interest, extracted from their distinct features. We design a specific analysis pipeline for art events and we show it at work on real case studies.

3 Approach

The approach presented in this paper defines a specific KDD process that comprises some data enrichment and preprocessing steps, followed by data mining phases which lead to significant knowledge extraction results in our scenarios. We propose the pipeline reported in Figure 1: first we extract all the required data from the social media platforms; in the next phase, we transform and enrich the data to proper formats for the subsequent analysis and then store it; and finally, data analysis techniques are applied on the clean, enriched and preprocessed data. The next sections describe each phase of the process in detail.

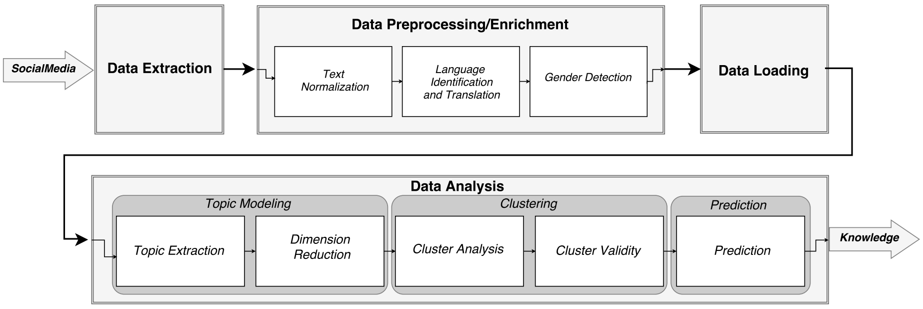


Fig. 1. Content analysis pipeline for art and culture events

3.1 Data Extraction

In this phase raw data is extracted by addressing the social network API with the appropriate query, which is able to extract information on the event of interest. We concentrate on Twitter as a good representative of social content in the context of live events and participation. Therefore, we exploit the Twitter API for data extraction [24].

3.2 Data Preprocessing

Since the collected raw data is incomplete and inconsistent, as described next, we need to apply preprocessing techniques to prepare an appropriate dataset which can be used for next analyses and experiments. The preprocessing phase consists of three main steps to be followed:

- **Text Normalization:** Textual properties (especially in social media) include a great deal of non-standard characters, punctuation, symbols, stop words, etc. that must be omitted for making the data clean and standard. Furthermore, it is essential to reduce derived words to their word stem or root form.
- **Language Identification and Translation:** Unsurprisingly, social media users do not always tweet in English, so having text in different languages are not unexpected. The majority of research works only focus on English contents, and thus the importance of language as a demographic feature

Table 1. Data Schema

Entity	Attribute	Description
Tweet	Id	The representation of the unique identifier for this tweet.
	Username	The user who posted this tweet.
	Text	The actual UTF-8 text of the status update.
	Date	Date and time when this tweet was created.
	Retweets	Number of times this tweet has been retweeted.
	Favorites	Indicates how many times this tweet has been liked
	Mentions	The users who are mentioned in this tweet.
	Hashtags	Represents hashtags which have been parsed out of this tweet text.
	Geo	Represents the geographic location of this tweet
User	Id	The representation of the unique identifier for this user.
	Username	The unique name of this user.
	Full name	The name of this user, as theyve defined it.
	Tweets	The users most recent (20) tweets.
	Follower count	The number of followers this user currently has.
	Following count	The number of users this user is following.
	Status count	The number of tweets issued by this user.
	Listed count	The number of public lists that this user is a member of.
	Favorite count	The number of tweets this user has favorited.
	Bio	The user-defined UTF-8 string describing their account.
	Hashtags	The hashtags included in this users most recent (20) tweets.
	Mentions	The users who are mentioned in this users most recent (20) tweets.
	Location	The user-defined location for their profile.
	Language	The users self-declared user interface language.
	Gender	Represents gender of user (added after data preprocessing).
	Lists	Names of the lists that user is member of.

is overlooked. Hence, with the aim of making data more coherent and unambiguous, and for expanding the coverage of the approach to world-wide scenarios, we apply language detection and translation into English, for homogenisation.

- **Gender Detection:** Twitter does not provide users' gender in their objects. Since we consider this a crucial demographic feature, we enrich the data with gender information.
- **Data Loading:** In this phase the clean and enriched data is stored in appropriate format for large scale analysis (CSV file).

3.3 Data Analysis Overview

In order to avoid manual tagging of data, which would be costly and not scalable across multiple experiment or usage scenarios, we opt for unsupervised techniques, namely clustering and topic modeling.

Since we want to profile users based on the texts they share on Twitter, first we need to create a Document-Term Matrix (DTM) which discovers the importance (frequency) of terms that occur in a collection of documents. It is noteworthy that, from now on, our "document" of interest is the social network user. Therefore, in practice a document corresponds to each user's textual feature, namely: personal *biography*, *hashtags* used in the tweets, text of the *tweets* posted, and *Twitter lists* the user belongs to (see Table 1). Therefore, each entry of DTM contains the frequency of each term occurred in each document.

As one can easily understand, this matrix is very big and extremely sparse. With the objective to get a more high-level and understandable sense of the documents, we apply topic extraction by means of Latent Dirichlet Allocation (LDA) on the matrix. The output of LDA is also a matrix that assigns a probability to each pair of document and extracted topic: in practice, we obtain a probability of a document (i.e., user) to be interested in a given topic. We then use this LDA output for clustering users.

We also define a prediction phase, where we use a classification method (specifically, Decision Trees), to create a model that can anticipate whether a newcomer user might be interested in the event, and can predict the topic(s) of interest for that user.

3.4 Topic Modeling

Topic models can help to organize and offer insights to understand large collections of unstructured text bodies [20]. They allow the probabilistic modeling of term frequency occurrences in documents. The fitted model can be used to estimate the similarity between documents as well as between a set of specified keywords using an additional layer of latent variables which are referred to as topics [14]. The input data for topic models is a document-term matrix (DTM). The rows in this matrix correspond to the documents and the columns to the terms. The entry m_{ij} indicates how often the j^{th} term occurred in the i^{th} document.

In this study, topic modeling phase consists of two steps:

- **Topic Extraction:** to discover the abstract topics that occur in the collection of our documents, we apply a topic model such as *Latent Dirichlet Allocation (LDA)* which benefits from Gibbs sampling algorithm [12]. For fitting the LDA model to a given document-term matrix, the number of topics needs to be fixed a-priori. Because the number of topics is in general not known, models with several different numbers of topics are fitted and the optimal number is determined in a data-driven way [14]. Maximum values of Deveaud et al. (2014) method and minimum values of Cao et al. (2009) estimation are considered optimal and are used in this study to identify the number of topics in LDA. The output of this model is a *topic probability matrix* that contains the probability of each topic associated to each document. In practice, this tells us the probability that a given user is interested in a given topic.
- **Dimension Reduction:** Since the extracted topics from LDA are possibly correlated, it is suggested to employ *Principal Component Analysis (PCA)* to convert them to a set of values of linearly uncorrelated topics. This transformation of data to a lower dimensional feature space not only reduces the time and storage required but also makes the data visualization and interpretation easier.

3.5 Clustering

Clustering aims to organize a collection of data items into clusters, such that items within a cluster are more similar to each other than they are to items in the other clusters [13]. In this work, this phase of the pipeline is divided into two steps:

- **Cluster Analysis:** In order to profile social media users based on the texts they share about a specific event, different cluster algorithms namely *K-means*, *Hierarchical*, and *DBSCAN* are used and compared, in order to select the best option in our specific setting.
- **Cluster Validity:** When cluster analysis was performed, it's crucial to evaluate how good the resulting clusters are. The evaluation indices, that are applied to judge various aspects of cluster validity are traditionally classified into three types: unsupervised (internal), supervised (external), and relative [23].

  In this study, *Silhouette Coefficient* and *Dunn's Index* as internal indices and *Entropy* as an external criterion are selected in order to evaluate and compare the different aspects of clustering results.

3.6 Prediction of User Interest

In order to guide event planning professionals to market, plan and implement their events more effectively, we propose to predict the category or the interest area of potential new users who might be involved in the similar cultural or art events in the future. Accordingly, beside the outlined unsupervised techniques that were employed to profile users, we opt for a supervised learning algorithm namely *decision tree*, which builds a classification model, for prediction of new users' interests, based on the user categories that we obtained from clustering. Decision tree learning is a typical inductive algorithm based on instance, which focus on classification rules displaying as decision trees inferred from a group of disorder and irregular instance [5].

To build the decision tree, first our dataset should be divided into training and test sets, with training set used to build the model and test set used to validate it. Since our target is to predict the interest domain of new users in terms of *user category*, the tree will be fed with the training dataset in which the category of users has been attached to and used as the target feature.

In additiion, the input variables comprise several features of the user, namely: the topic probabilities defined for each textual features of the user (coming from the *topic probabilities matrix* generated by the LDA analysis), namely the biography, hashtags, tweets and lists; gender, language, number of followings and followers and number of tweets. Consequently, the decision tree generates a set of prediction rules that determine new users' interest areas, regarding a cultural or art event, based on the values of the features.

4 Implementation

In the preprocessing phase, we use the Yandex API [27] for language identification and translation of stems into English, and the NamSor API [19] for gender detection. In the preliminary phases we store the data in a relational database for fast sequential preprocessing, and then we generate CSV files to be fed to the analysis phases.

All analyses, statistics, evaluations and results representations are done in R, a flexible statistical programming language and environment that is open source and freely available for all mainstream operating systems [17].[1]

5 Case Study

For sixteen days, from June 18 through July 3 2016, Lake Iseo in Italy was reimagined by the world-renowned artists Christo and Jeanne-Claude.[2] More than 100,000 square meters of shimmering yellow fabric, carried by a modular floating dock system of 220,000 high-density polyethylene cubes, undulated with

[1] We use `tm` package for all the text-mining methods for importing data, corpus handling, preprocessing and creation of document-term matrices; `topicmodels` package for LDA; `cluster`, `fpc`, `dbscan` and `NbClust` for clustering and cluster validation; `ggplot2`, `rgl`, `ggmap`, `wordcloud` and `RColorBrewer` packages for interactive graphics; `rpart`, `rpart.plot` and `party` packages for decision tree modeling and prediction.

[2] `http://christojeanneclaude.net/projects/the-floating-piers`

the movement of the waves as The Floating Piers rose just above the surface of the water. Visitors were able to experience the work of art by walking on it from Sulzano to Monte Isola and to the island of San Paolo, which was framed by The Floating Piers [4] (see Figure 2[3]). More than 1.5 million people visited the installation in those 2 weeks.

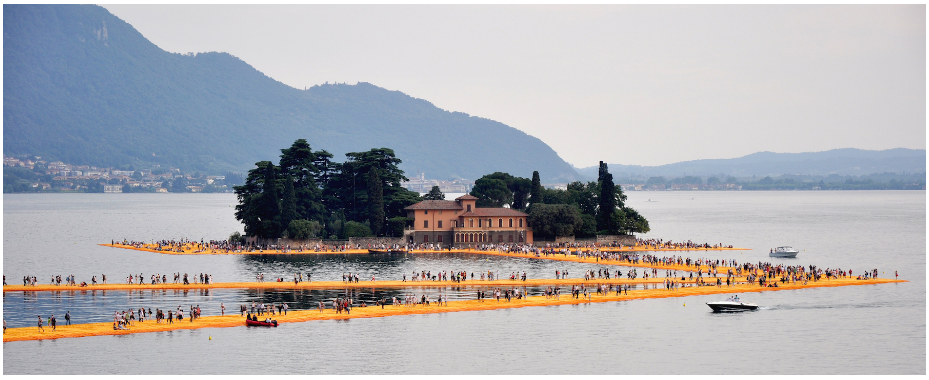


Fig. 2. The Floating Piers by Christo and Jeanne-Claude

We use this artistic event as a use case for our method. We extracted the social media content relevant to the event and we applied the analysis pipeline over it. The datasets were obtained from Twitter, during a time period from June 10th to July 30th 2016 and contain 14,062 tweets and 23,916 users. Figure 3 represents the total number of tweets, retweets, favorites and engaged users (per day), one week before the event starts until the end of July. As one can see, users tend to tweet about the installation at the early days of the event while the engagement of the users dramatically decreases afterwards.

According to the statistics, unlike Instagram users, most Twitter users are not willing to specify the location of their published tweets. Therefore, we also extracted Instagram posts (30,256 posts and 94,666 users) related to the event during the same time span and displayed the density of these 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.

6 Results and Discussion

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

[3] Photo Credits: Harald Bischoff, *Christo's "The Floating Piers", lake of Iseo, Italy 2016.* License: Creative Commons Attribution-Share Alike 3.0 Unported.

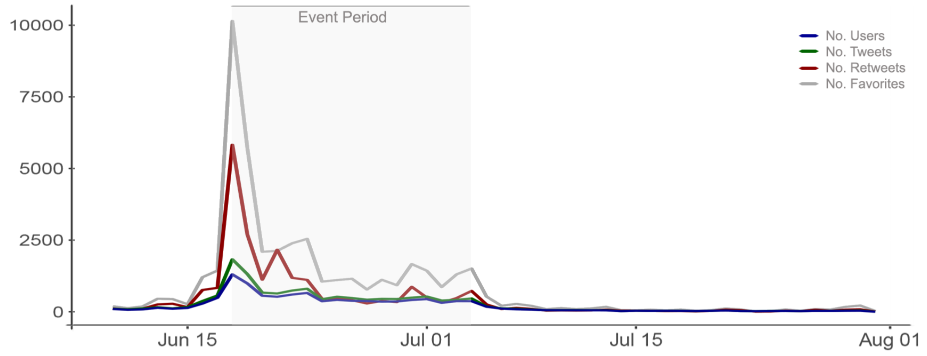
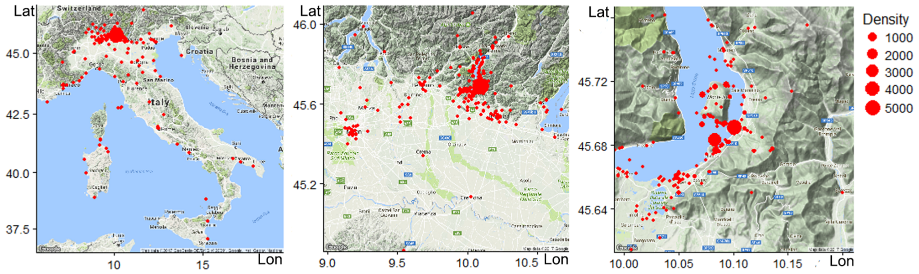


Fig. 3. Tweets, Retweets, Favorites and Users Timeline



(a) Italy (b) Lombardy Region (c) Iseo Lake

Fig. 4. Density of Instagram posts in different coordinates

6.1 User Clustering

As discussed earlier, we apply and compare different clustering algorithms, namely K-means, Hierarchical and DBSCAN. Each of them is separately applied on our data collections. Each collection consists of documents that correspond to each textual property of users including bio, hashtags, tweets and lists.

Subsequently, to achieve the most accurate results, different cluster validity measures are employed. Among the existing validation metrics, Silhouette width, Dunn index and Entropy are selected to evaluate the clustering results. Silhouette width and Dunn index combine measures of compactness and separation of the clusters. Thus, algorithms that produce clusters with high Dunn index and high Silhouette width are more desirable. On the other hand, Entropy is a metric that is a measure of the amount of disorder in a vector. So, smaller values of entropy indicate less disorder in a clustering, which means a better clustering [7].

Table 2 represents these measures values for each algorithm that was applied on each feature collection. It should be noted that, the number of clusters in each experiment is either determined by the clustering algorithm itself like DBSCAN or calculated through different methods like Elbow for K-means. According to this table, Hierarchical clustering (with three clusters) can be considered as the best algorithm which has produced more pleasant results compared to the other two. Furthermore, among all four examined textual features, users' biography (Bio) performs best. Consequently, as table suggests (bold numbers), from now on we only focus on hierarchical clustering performed on users' biography.

Table 2. Evaluation of clustering results

		Features			
Algorithms	Indices	*Bio*	*Hashtags*	*Tweets*	*Lists*
K-Means	No.Clusters	3	3	4	3
	Silhouette	0.457	0.427	0.425	0.535
	Dunn	0.047	0.003	0.001	0.001
	Entropy	1.070	0.799	0.725	0.736
DBSCAN	No.Clusters	5	4	5	5
	Silhouette	0.051	0.329	0.106	0.270
	Dunn	0.020	0.008	0.001	0.001
	Entropy	1.207	0.053	0.716	0.394
Hierarchical	No.Clusters	3	3	3	3
	Silhouette	**0.595**	0.520	0.506	0.331
	Dunn	**0.050**	0.006	0.001	0.001
	Entropy	**0.015**	0.377	0.725	0.905

6.2 Applying Topic Modeling

As mentioned earlier, in this study, the input data for clustering models is a topic probability matrix that contains the probability of each topic associated to each document (user). This matrix is generated after applying LDA on document-term matrix, in which each row is a user' biography and each column is a term. As indicated in section 3.4, Deveaud et al. (2014) and Cao et al. (2009) can help to determine the number of topics before LDA is applied. Accordingly, the optimum values of these metrics offer 6 as the number of topics for LDA. Having all required parameters set, LDA is applied and returns the topic probability matrix along with the top terms of each extracted topic which are presented in a word network in Figure 5.

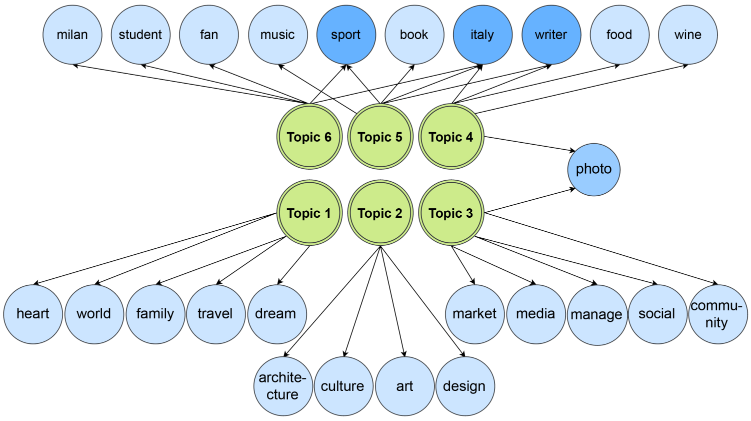


Fig. 5. Word network representation of top terms in each topic

By investigating through these terms, it seems that the extracted topics are correlated and need to be transformed to a lower-dimensional set, supplied by PCA procedure. The result of applying PCA on topics is represented in Table 3.

Table 3. PCA quantitative results

	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6
Standard Deviation	1.1188726	1.0683576	0.7789327	0.7789327	0.44049766	1.290478e-07
Proportion of Variance	0.4172919	0.3804627	0.2022454	0.1022776	0.03880764	2.775558e-15
Cumulative Proportion	**0.4172919**	**0.7977546**	**1.0000000**	1.000000e+00	1.000000e+00	1.000000e+00

As the table suggests, we only consider the first three principal components (topics) where cumulative proportion passes 95 percent threshold. Consequently, in clustering phase, we will perform the hierarchical algorithm on topic probability matrix, exploiting only these three PCA-selected topics.

6.3 Cluster Hierarchy

As indicated in previous sections, hierarchical algorithm returns more acceptable results. This algorithm's output is a dendrogram which is illustrated in Figure 6.

Unlike K-means algorithm, hierarchical algorithm does not require the optimal number of clusters to be defined at the beginning. In this clustering algorithm clusters are defined by cutting branches off the dendrogram. To determine the cutting section, various methods can be used. We used a convention which represents that a dendrogram can be cut where the difference is most significant.

To extract better insights over the situation, we report in Figure 6 the three main clusters drawn in different colors. Each leaf in this tree represents a Twitter user engaged in the Floating Piers event through tweeting, retweeting or liking a post. According to this result, it can be concluded that nearly 60 percent of users lies in first cluster (green), over 35 percent in second (blue) and the rest (about 5 percent) in the third cluster (red).

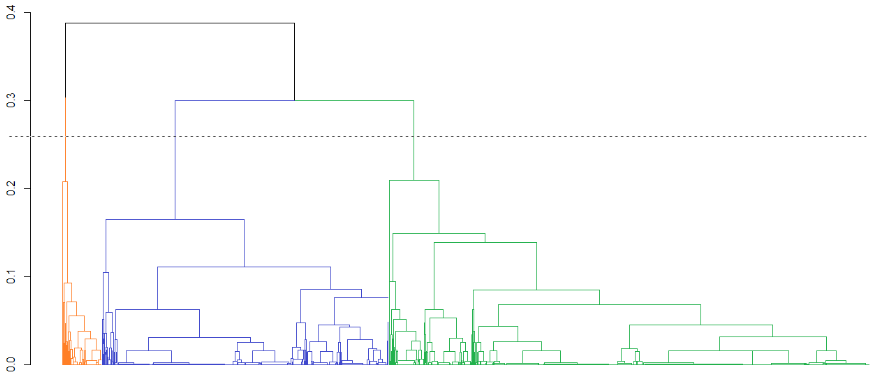


Fig. 6. Dendrogram representation of Twitter users

6.4 Cluster Labeling/User Profiling

Having all the user objects in each cluster, we are able to label the obtained clusters or in other words to identify the categories of users. The five most frequent words, that users published about the event, along with the frequency of each word are indicated in Table 4. In this table, Cluster 1 refers to the biggest cluster (green), Cluster 2 refers to the second biggest cluster (blue) and Cluster 3 refers to the smallest cluster (red). It can be seen that the most frequent words in each cluster convey specific meanings. People in first cluster mostly talk about

Travel introducing themselves in their Twitter bio. People in second cluster are Art lovers and people in third cluster state their positions as Technology fans and social media marketing addicted. Henceforth, we call the users in first, second and third cluster Travel Lovers, Art Lovers and Tech Lovers respectively.

Table 4. The five most frequent words and their frequency in each cluster

Cluster 1 (Travel)		Cluster 2 (Art)		Cluster 3 (Tech)	
Word	Freq.	Word	Freq.	Word	Freq.
travel	1070	art	962	social	677
italian	832	design	811	market	641
passion	797	culture	749	media	618
lover	632	photograph	598	manag	502
food	614	artist	537	founder	435

To depict a weighted list of the words that are used in users bio in each cluster, we employ word networks and word clouds, which are visual representations of textual data. The word networks and word clouds related to users' bio in each cluster are illustrated in Figure 7 and Figure 8 respectively.

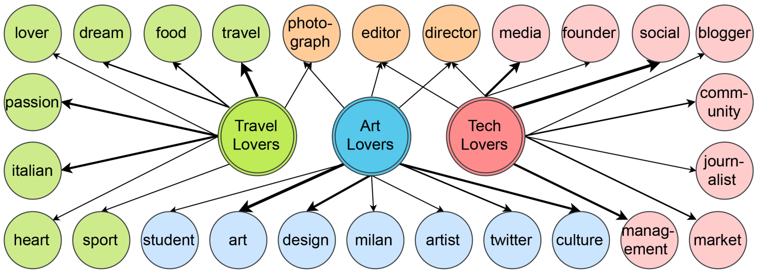
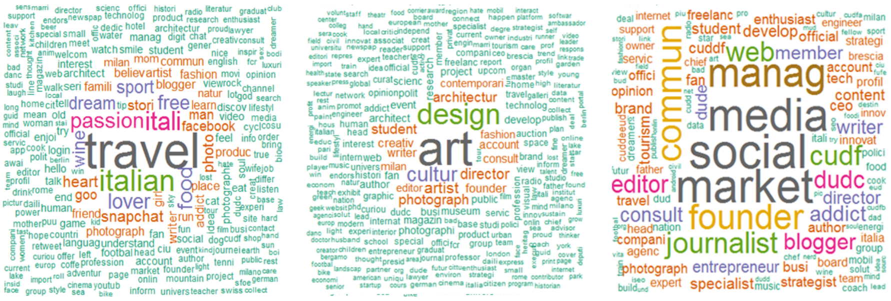


Fig. 7. Word network representation of top terms in each cluster - the thickness of the connections is proportional to the frequency of words in clusters

6.5 Demographic Analysis - Language

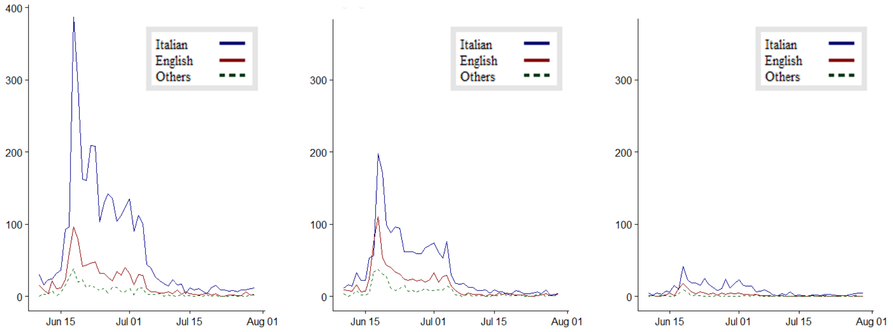
We can use demographic features like language and gender which help to compare users in three clusters. As Figure 9 shows Italian is the most common language of users in three clusters while second place belongs to English, followed by the sum of all the other languages (French, Dutch, etc.). As one can see, the flows of languages follow the flow of tweets in all three clusters and have a peak on



(a) Travel Lovers (b) Art Lovers (c) Tech Lovers

Fig. 8. Word cloud for each cluster

the opening day of the event. The bias towards Italian is particularly evident in the travel lovers cluster, while it's less strong in the art lovers. This suggests that travelers visiting the event are mostly Italians, while people coming from abroad are not generic tourists, but more specifically art lovers, which come on purpose for the event.

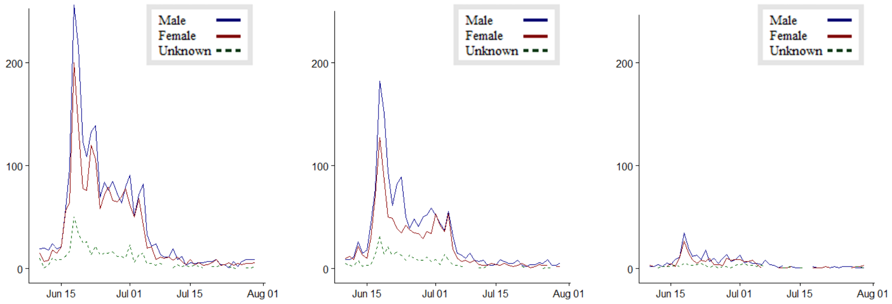


(a) Travel Lovers (b) Art Lovers (c) Tech Lovers

Fig. 9. Time series of posts by language for each cluster

6.6 Demographic Analysis - Gender

Figure 10 demonstrates that the number of males who got involved in the Floating Piers overweighs the number of females but the difference is not substantial and can be overlooked. In addition, since Travel lovers are the highest majority, the number of males and females are the highest in this category. The presence of males is slightly higher in art lovers.



(a) Travel Lovers (b) Art Lovers (c) Tech Lovers

Fig. 10. Time series of posts by gender for each cluster

6.7 Prediction of Interests of New Users

As mentioned in section 3.6, we suggested to employ decision tree to predict the possible interests of the potential future users, based on the categories that were acquired from clustering the current users. There are two competing concerns: with less training data, our parameter estimates have greater variance. With less testing data, our performance statistic will have greater variance. Thus, we divide our dataset into training and testing sets with the ratio of 80:20, such that neither variance is too high. Figure 11 can give an intuition of how the decision tree creates the prediction rules.

In addition, the extracted rules from the tree are formulated as follows:

```
- Rule 1: if (0.36 < Bio_score < 0.37 OR Bio_score < 0.35) then new user
                        is interested in event and is a Travel Lover

- Rule 2: if (0.35 < Bio_score < 0.36 AND Status_count > 14.5) OR
             (Bio_score > 0.37 AND language != Italian) then new user
                        is interested in event and is an Art Lover

- Rule 3: if (Bio_score > 0.37 AND Language = Italian) then new user
                        is interested in event and is a Tech Lover

- Otherwise: then new user is NOT interested in event
```

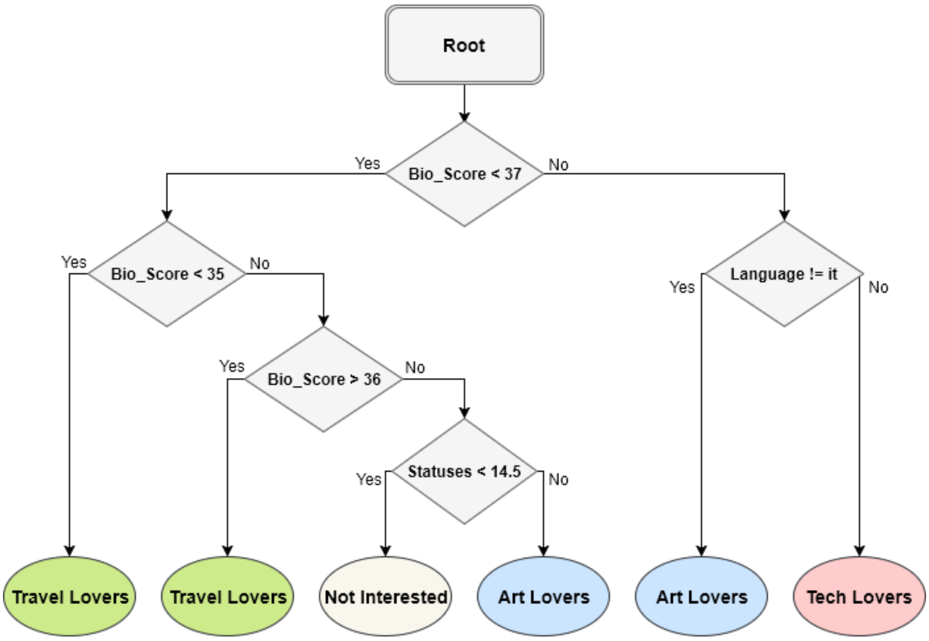


Fig. 11. The Decision Tree Representation

where the only relevant features identified by the decision tree are: the `Bio_score`, representing the topic probability for the biography feature; the `Status_count`, representing the total number of tweets of the user; and the `Language`.

In order to evaluate the decision tree, we use the test dataset that determines an accuracy of 62 percent. Now for new users, we can simply use the above rules and specify their categories (Travel, Art or Tech) or identify them as not interested in a similar event in the future.

Notice that the power of the solution, considering the obtained rules, is that of being able of classifying a user as interested and as engaged in travel, art or technology essentially looking at its biography.

7 Conclusion and Future Work

In this study, we proposed a complex approach that addresses user profiling and user interest prediction regarding arts and cultural events on social media. This approach is equipped with a preprocessing step that enriches user data in terms of language and gender. The outcomes of this research can help event organizers to decide what categories of users they are dealing with and have a clear understanding about the characteristics of users who are more likely to be attracted by the similar events in the future.

We used The Floating Piers event as a case study to show how the proposed approach works with the real life scenarios. We clustered users based on their interests in three main categories and then described and compared the behavior and properties of users in each cluster. In addition, using decision tree modeling resulted in a set of rules that predicts the interest domain of future users.

However, since the current study merely addresses the text content that users share on social media, it can go further with considering other types of media namely photo in other social network platforms such as Instagram, Facebook, Google+, Flickr, Foursquare, etc. that might result in a clearer and wider picture of the characteristics and behaviour of users with respect to cultural and artistic events. Last but not least, applying other techniques like semantic analysis, image processing and network analysis can also help us to improve the accuracy and coverage of the results.

References

1. Arabghalizi, T., Rahdari, B.: Event-based User Profiling in Social Media Using Data Mining Approaches. Master's thesis, Politecnico di Milano (April 2017)
2. 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
3. Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M., Vesci, G.: Choosing the right crowd: expert finding in social networks. In: Joint 2013 EDBT/ICDT Conferences, EDBT '13 Proceedings, Genoa, Italy, March 18-22, 2013. pp. 637–648 (2013), http://doi.acm.org/10.1145/2452376.2452451
4. Christo: The floating piers (2016), http://christojeanneclaude.net/projects/the-floating-piers
5. Dai, Q.y., Zhang, C.p., Wu, H.: Research of decision tree classification algorithm in data mining. International Journal of Database Theory and Application 9, 1–8 (2016)
6. DIAO, Q.: Event identification and analysis on Twitter. Ph.D. thesis, Singapore Management University (2015)
7. Dziopa, T.: Clustering validity indices evaluation with regard to semantic homogeneity. In: Position Papers of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016, Gdańsk, Poland, September 11-14, 2016. pp. 3–9 (2016), https://doi.org/10.15439/2016F371
8. 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
9. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: Advances in knowledge discovery and data mining. chap. From Data Mining to Knowledge Discovery: An Overview, pp. 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA (1996), http://dl.acm.org/citation.cfm?id=257938.257942
10. Friedemann, V.: Clustering a Customer Base Using Twitter Data. Tech. rep., stanford university department of computer science (10 2015)
11. Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., Gummadi, K.: Cognos: Crowdsourcing search for topic experts in microblogs. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 575–590. SIGIR '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2348283.2348361

12. Griffiths, T.: Gibbs sampling in the generative model of latent dirichlet allocation. Tech. rep. (2002)
13. Grira, N., Crucianu, M., Boujemaa, N.: Unsupervised and semi-supervised clustering: a brief survey. In: in A Review of Machine Learning Techniques for Processing Multimedia Content, Report of the MUSCLE European Network of Excellence (FP6 (2004)
14. Grun, B., Hornik, K.: Topicmodels: An r package for fitting topic models. Journal of Statistical Software 40(13), 1–30 (2011)
15. Hu, Y.: Event Analytics on Social Media: Challenges and Solutions. Ph.D. thesis, Arizona State University (2014)
16. Kanoje, S., Girase, S., Mukhopadhyay, D.: User profiling trends, techniques and applications. International Journal of Advance Foundation and Research in Computer (IJAFRC) 1, 119–125 (2014)
17. Kelley, K., Lai, K., Wu, P.J.: Using r for data analysis: A best practice for research. In: Osborne, J. (ed.) Best Practices in Quantitative Methods. SAGE publishing (2008)
18. Kisilevich, S., Krstajic, M., Keim, D., Andrienko, N., Andrienko, G.: Event-based analysis of people's activities and behavior using flickr and panoramio geotagged photo collections. In: 2010 14th International Conference Information Visualisation. pp. 289–296 (July 2010)
19. NamSorSAS: Namsor api, `http://www.namsor.com`
20. Patil, M., Kankal, S.: Topic digging over asynchronous text sequences. International Journal Of Engineering And Computer Science 5, 19548–19551 (2016)
21. Prangnawarat, N., Hulpus, I., Hayes, C.: Event analysis in social media using clustering of heterogeneous information networks. In: The Twenty-Eighth International Flairs Conference (2015)
22. Singh, K., Shakya, H.K., Biswas, B.: Clustering of people in social network based on textual similarity. Perspectives in Science 8, 570 – 573 (2016), `http://www.sciencedirect.com/science/article/pii/S2213020916301628`, recent Trends in Engineering and Material Sciences
23. Tan, P.N., Steinbach, M., Kumar, V.: Cluster analysis: Basic concepts and algorithms. In: Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc. (2005)
24. Twitter: Api overview (2017), `https://dev.twitter.com/overview/api`
25. Van Craenendonck, T., Blockeel, H.: Using internal validity measures to compare clustering algorithms. In: Benelearn 2015 Poster presentations (online). pp. 1–8 (2015)
26. Xiao, J., Zhang, Y., Jia, X., Li, T.: Measuring similarity of interests for clustering web-users. In: Proceedings of the 12th Australasian Database Conference. pp. 107–114. ADC '01, IEEE Computer Society, Washington, DC, USA (2001), `http://dl.acm.org/citation.cfm?id=545538.545551`
27. Yandex: Translate api (2014-2017), `https://tech.yandex.com/translate`
28. Zhou, G., Ding, H., Zhou, G., Zhang, W.: A user clustering algorithm considering user's interest-offset. In: International Conference on Cyberspace Technology (CCT 2013). pp. 62–67 (Nov 2013)