A Predictive Method to Improve the Effectiveness of Twitter Communication in a Cultural Heritage Scenario

MARCO FURINI, University of Modena and Reggio Emilia
FEDERICA MANDREOLI, University of Modena and Reggio Emilia
RICCARDO MARTOGLIA, University of Modena and Reggio Emilia
MANUELA MONTANGERO, University of Modena and Reggio Emilia

Museums are embracing social technologies in the attempt to broaden their audience and to engage people. Although social communication seems an easy task, media managers know how hard it is to reach millions of people with a simple message. Indeed, millions of posts are competing every day to get visibility in terms of likes and shares and very little research focused on museums communication to identify best practices. In this paper, we focus on Twitter and we propose a novel method that exploits interpretable machine learning techniques to: (a) predict whether a tweet will likely be appreciated by Twitter users or not; (b) present simple suggestions that will help enhancing the message and increasing the probability of its success. Using a real-world dataset of around 40,000 tweets written by 23 world famous museums, we show that our proposed method allows identifying tweet features that are more likely to influence the tweet success.

CCS Concepts: • **Human-centered computing → Social media**; *Empirical studies in collaborative and social computing*.

Additional Key Words and Phrases: Twitter, machine learning, prediction

ACM Reference Format:
Marco Furini, Federica Mandreoli, Riccardo Martoglia, and Manuela Montangero. 2021. A Predictive Method to Improve the Effectiveness of Twitter Communication in a Cultural Heritage Scenario. *ACM J. Comput. Cult. Herit.* 1, 1, Article 1 (January 2021), 19 pages. https://doi.org/10.1145/3470786

1 INTRODUCTION

Museums need a communication strategy to broaden their visitor base, which is mostly made up of adults. In this perspective, social media represent a great opportunity as they are used by young users who look at and experience culture differently than in the past [20]. It is not surprising that the key point of the agenda of many museum media managers is the design of a communication strategy able to engage users in the hope of getting them interested in art and turning them into (real) visitors.

Although everyone agrees with the enormous potential of the social scenario, the way these social platforms are used is very disappointing. Instead of trying to involve users, the communication is usually limited to naïve marketing and outreach strategies that result in one-way communication, events listings, reminders, promotions

Authors' addresses: Marco Furini, University of Modena and Reggio Emilia, Dip. Comunicazione ed Economia, Reggio Emilia, Italy, marco.furini@unimore.it; Federica Mandreoli, University of Modena and Reggio Emilia, Dip. Scienze Fisiche, Informatiche e Matematiche, Modena, Italy, federica.mandreoli@unimore.it; Riccardo Martoglia, University of Modena and Reggio Emilia, Dip. Scienze Fisiche, Informatiche e Matematiche, Modena, Italy, riccardo.martoglia@unimore.it; Manuela Montangero, University of Modena and Reggio Emilia, Dip. Scienze Fisiche, Informatiche e Matematiche, Modena, Italy, manuela.montangero@unimore.it.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
XXXX-XXXX/2021/1-ART1 $15.00
https://doi.org/10.1145/3470786

and announcements [8, 9, 24]. Very recently, museums are uploading images to Facebook pages or to Instagram accounts to rise interest towards art and culture [16] and measure the engagement by counting the number of followers they have. This approach has two weaknesses: the first is that a simple list of images is not enough to engage users [14] and the second is that the number of followers is a simple data that does not turn into useful information (i.e., the accounts you follow can be put into lists or can be silenced, not to mention that the followers could be bots [22]).

The scenario is very complex and little research specific to museum communities has been performed to determine best practices on social media platforms [2]. For example, a museum like the Guggenheim in New York has millions of followers, while others like the Art Institute of Chicago have far fewer. What brings people to follow a museum? There are museums that post more than one message per day and museums that rarely post messages. What is the best approach, if there is one? Some messages use images, icons, emoticons, long texts. Is there a reason behind these choices? Why do some messages go viral and others don't? Are there any criteria that influence the virality of messages? These unanswered research questions motivate us to broaden the theoretical framework proposed in current studies by defining a method for implementing a museum-tailored social media communication strategy. In particular, we focus on the Twitter platform and we propose a method that properly combines a *tweet classifier* and a *suggestion mechanism* into a strategy to make on-the-fly tweet success prediction. By leveraging on tweet content and on cutting-edge interpretable machine learning techniques, our proposal gives insights to media manager writing tweets. For instance, they might receive suggestions like: *"You should drop at least two or three URLSs"*, *"You should use an image"*.

Our proposed method is based on the analysis of a dataset composed of around 40,000 tweets posted by 23 world famous museums (1,700 tweets per museum on average). First, we identify the successful tweets (i.e., the ones that engaged most users) and we investigate which machine learning technique (among K-nearest neighbor, Decision Tree, Random Forest and Ada Boost) is best suited to the scenario under consideration. Then, we use the best suited machine learning technique to learn the features (e.g., number of hashtags, number of media contents, words, hashtags and mentions) that most influence the success of a tweet. The obtained bag of knowledge is then used to predict the success of a tweet at writing time. The obtained results show that our proposed method predict the success of a tweet (discriminating between "good" and "bad" ones) with an average accuracy higher than 85%.

With respect to our previous preliminary studies [9–11, 19], the method (and the associated dashboard UI) presented in this paper builds on them and for the first time is: (a) based on interpretable machine learning techniques; (b) applied to a large dataset with a diversified feature set (including both countable and on-off content, context and user features); (c) discussed and tested in depth through a comprehensive evaluation which also includes the assessment of the suggestion mechanism and the detailed model interpretation discussion on how each feature impacts the model.

The contribution of this study is twofold. From the theoretical point of view, it identifies a method to make predictions in a scenario, that of museums, rarely taken into consideration in literature: it is a first step towards a communication strategy capable of restoring users' interest in museums. From a practical point of view, our method allows museum media managers to improve users' engagement. Indeed, our method is easy to implement (as shown through our developed dashboard) and is democratic (i.e., it can be used by museums with rich communication plan budget, and by museums with poor or no communication budgets).

Both the complete dataset and the analysis code used in this paper are publicly available for download (links in Section 4).

The rest of the paper is organized as follows: after an overview of the state of the art (Section 2), in Section 3 we describe our method: how we prepared our dataset and the experiments on predictive analysis for tweet classification (Section 3.1), the suggestion mechanism (Section 3.2) and an overview of the dashboard that implements our method and its functioning (Section 3.3). Section 4 presents the results of the experiments we

performed to assess the validity of our method. Finally, Section 5 provides a detailed discussion on the model interpretation results and Section 6 concludes the paper.

2 BACKGROUND AND RELATED WORK

Twitter is among the most well-known and used social network platforms. On average, over 300 million Twitter users are active each month and post more than 500 million tweets every day. In addition to its institutional role of displaying news and reports, Twitter is mainly a global place where people freely share their feelings, tell their location, express what they are doing and exchange opinions on a wide range of topics, from politics to products [26]. Users of Twitter use different means to enrich their messages, from multimedia to hashtags employed to add context and metadata [21].

In particular, thanks to its widespread diffusion, Twitter can represent a valuable media communication channel for museum curators to engage with visitors and promote exhibitions. This phenomenon has been subject to data analytic studies for some years. A quantitative and qualitative approach to the analysis of tweets posted during the MuseumWeek event organized by Twitter is proposed in [6]. The proposed data analysis framework first categorizes the social accounts into two main types, institutional accounts and personal accounts, then quantifies the attendance and, finally, automatically classifies tweets according to four different categories. The work [5] moves a step forward by aiming to describe how cultural organizations could exploit Twitter to reach potential visitors of exhibitions. The proposed approach essentially consists in identifying a set of Key Performance Indicators (KPIs) for quantitative estimation of Cultural Heritage Sensitivity as expressed by social network users in a given time period and geographic region. Twitter performances of the top 60 European museums are analysed in [24]. The analysis is performed by using Descriptive Analytics, Principal Components Analysis and correlation analysis and concerns networking, twitting activities, times that the account is active, and involvement. Analysis results reveal that museum performances are partially in accordance with popularity and ranking and allow the identification of a group of more active museums. [17] aims to understand more about how a museum is employing a level of engagement by seeking to answer two broad questions about social media use concerning the types of tweets and the kinds of activities the museum performs to engage their visitors. The analysis focuses on 48 museums across the US and finds that a large number of museums in the sample focused on original content in their Twitter feed. This suggests that Twitter use is a core business function with dedicated staff time for uploading and attending to postings.

In line with the papers above, this paper focuses on art museums. However, the goal of this work is not a coarse grain analysis of the use of social media by museums. Rather, we aim to a fine grain analysis of the features that characterize tweets of museum accounts and their use in predicting tweet influence of a given account in comparison with the competitors of the related museum. The findings of these papers about the different approaches the analysed museums adopt when using their Twitter accounts and the relationship with the museum popularity are therefore essential for our goals.

Different approaches have been proposed in the literature to analyze and predict tweet influence, a.k.a. popularity. Prediction is in most cases dealt with through machine learning approaches. For instance, paper [27] focused on news agencies accounts on Twitter and studies the propagation characteristics of news on Twitter as a backbone of a Twitter news popularity prediction model. The stochastic model is able to predict the final number of retweets of a news tweet. In this study, they also found that the negative sentiment of news has some correlation with tweet popularity while the positive sentiment does not have such obvious correlation. The work in [28] concerned with a popular micro-blogging website in China, Sina Weibo, and aims to discover content factors and contextual factors that affect the popularity of tweets. Four content-related factors are included in the study: topics, length, affective degree (number of emoticons and modal particles), and availability of supplementary information, while the considered context factors are users' degree of activeness, self-disclosure

degree, experience, and authoritativeness. Popularity of tweets was measured by considering the width of tweets distribution and depth of deliberation on tweets, i.e.,the number of comments tweets received. They found that the two factors are equally important to predict the first popularity measure, i.e. the tweets diffusion, but content outperforms context when predicting the second popularity measure, i.e., the number of comments tweets received. In [12], authors aimed to identify features for tweet popularity prediction that are both effective and effortless, i.e., easy to obtain or compute. From the experimental assessment, it followed that a relative small set of features, in particular temporal features, can achieve comparable performance to all features. The paper [29] approaches the problem of predicting the final number of reshares of a given post by adopting a different approach. Instead of relying on machine learning approaches, it proposes a statistical model that allows them to make accurate predictions through a simple and efficiently computable formula.

While all these papers focused on a notion of tweet popularity that is influenced by the network, our aim, instead, is mainly to study content features and their impact on popularity prediction. Moreover, the focus of our work is on art museums that, as to our knowledge, have never been studied in this context.

In our previous works [9, 10], we analyzed tweet contents extracted from museum accounts to investigate what features (e.g., images, hashtags, mentions, links, etc.) are worth using. The studies considered popular tweets posted by official museum accounts or ordinary people and analyzed them to derive insights about the tweet generation.

The impact of multimedia content on tweet popularity and life span was studied in [30]. The study showed that multimedia tweets dominate pure text both because they are more popular and because they survive longer. Moreover, sentiment analysis in Twitter is a field that has recently attracted research interest. An overview of the main algorithms that have been proposed for sentiment analysis in Twitter is provided in [13], whereas [25] investigated whether the community sentiment energy of a topic is related to the spreading popularity of the topic. Experiments on two communities found a linear correlation between the community sentiment energy and the real spreading popularity of topics. In our work [10], instead, we noticed that tweets originated by museums are typically neutral so in this paper we do not further investigate on this feature.

We conclude the related work analysis with some final notes on two further aspects that characterize and distinguish the research proposed in this paper. The first one is about the use of interpretable machine learning techniques. In general, there has been a recent emphasis on the need for explanations of ML systems [15] in many scenarios; one of the hottest ones is the medical setting, where clinicians desire to know the "reason" behind ML-based predictions (the specialized research trend is Interpretable Machine Learning in Healthcare [1]). Our aim in this research is to exploit past expertise in this field [7] to go beyond the black box nature of ML predictions also in the social media / cultural heritage context, where this has seldom been done.

The final note is on the use of dashboards in social media contexts. The concept of dashboard as a user interface providing at-a-glance views relevant to particular objectives is very common in the business sector and also in the context of social media and Twitter several examples exist. The Tweet activity dashboard[1] shows the user a detailed analysis of his/her activity on Twitter: for each sent message, the user is able to track impressions, engagements and engagement rate (impressions divided by engagements). IOM's Twitter dashboard[2] gathers all the Organization's global, regional special projects accounts. It provides a single repository for all the latest news about IOM's work all over the world and in particular, latest updates on migration. All these examples, however, are not focused on the cultural heritage / museums sector we consider and do not exploit intelligent data analytics features to give useful suggestions to users.

[1]https://business.twitter.com/it/analytics/tweet-activity-dashboard.html
[2]https://www.iom.int/dashboard/ (International Organization for Migration)

3 METHOD

The method we propose properly combines a *tweet classifier* and a *suggestion mechanism* into a strategy to predict tweet success on-the-fly.

- **Tweet classifier** (Section 3.1): our first goal is to classify messages to predict whether they will be successful or not, where success is measured by the number of retweets tweets receive. Indeed, a retweet is an intentional action taken by the user reading the tweet that implies a strong engagement to the tweet. We selected a set of tweet features that depend both on the content and the context of the tweet. Messages are then classified into GOOD (i.e., high success probability) and BAD (i.e., small success probability) by a classifier that has been trained with tweets sent by museums having a similar number of followers.
- **Suggestion mechanism** (Section 3.2): our second goal is to propose suggestions for enhancements to apply to a draft tweet message that has been classified as BAD (i.e., prediction is that it will not be successful). In general, suggestions should be easy to implement and should imply small but incisive changes in the message, otherwise the only sensible suggestion would be to rewrite the tweet from scratch. Therefore, we propose a mechanism that identifies the draft feature values that mostly contributed to the BAD classification and suggests new values according to GOOD tweets in the draft neighborhood.

From the practical point of view, we present a **Dashboard prototype** (Section 3.3) for mobile devices that implements the classifier and the suggestion mechanisms.

3.1 Tweet classifier

The task of the classifier is to predict whether a draft tweet message written by a museum media manager will be successful (GOOD) or not (BAD) once posted on the museum Twitter profile. Classification is done on a dataset of around 40,000 tweets posted by 23 world famous museums (see Section 4.1 for more details on the dataset) using supervised machine learning techniques trained on tweet feature vectors. The main aim of the approach is to study how content choices impact on popularity prediction and, thus, to offer easy-to-follow content suggestions to the media manager composing tweets so to make them more successful. For this reason, we consider commonly used features (see e.g. [3, 28]) and select those that can be easily understood by the user and extracted from the content of the tweet itself; these are complemented with a small number of additional features describing the context and the sender of the tweet and that might enhance the effectiveness of the classification process. The resulting feature set is composed of:

- *Content features:* these are the features that can be drawn form the content of the tweet itself. We further divide such features in the following two categories:
  - *Countable*: these features have a value ranging into different intervals. We take into consideration: the number of *hashtags* (i.e., words preceded by #) in the tweet, the number of *URLs* (i.e., links to external resources), the number of *images* (e.g., photos and graphical emoticons), the number of *mentions* (i.e., twitter accounts preceded by @), the *length* of the tweet;
  - *On-Off*: these features have binary values in $\{0, 1\}$. We observe whether the tweet has *exclamation marks*, *question marks*, *person names*, *place names*, *organization names*, *other names*. To compute the values of these features the Named Entity Recognition technique of the SpaCy library[3] is used. Moreover, we also take into consideration the tweet topic *density*: assuming that the involved topics correspond to the hashtags mentioned in the text, we define a tweet as dense of topics if the number of hashtags it contains is greater than a given threshold, set to 5 [28]. Finally, we observe the tweet sentiment that might be

[3] `http://spacy.io`

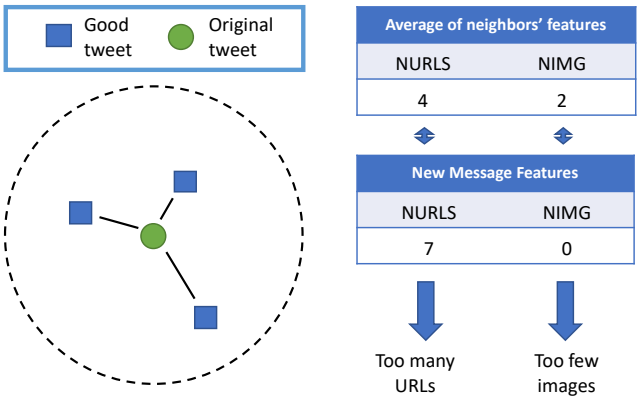


Fig. 1. Suggestion generation example.

present (positive or negative) or not (neutral) and has been computed using AFINN Sentiment Analysis library[4].

- *Context features*: these features are not drawn form the content of the tweet itself and might give a larger picture of the context in which the tweet was sent. Namely, we take into consideration the part of the day in which the tweet was sent (morning, afternoon, evening and night respectively from 5:00am to 11:59am, from 12:00pm to 5:59pm, from 6:00pm to 10:59pm and from 11pm to 4:59am), and a boolean feature indicating whether the tweet is a retweet or not.
- *User features:* these features are proper of the user that sent the tweet, and are the same for all the tweets of this user. Namely we consider the name of the museum and the number of followers of the user.

In order to train the machine learning model, tweet representations in the dataset are labelled GOOD or BAD according to the number of retweets (a retweet might be considered a metric to measure the user's engagement). The labelling process is museum dependent, i.e., the number of retweets to classify a tweet as GOOD or BAD is not the same for all the museums. For instance, a tweet with 5 retweets might be labelled as GOOD if posted by a small locally known museum, or might be labelled as BAD if posted by a museum with millions of followers. As we will detail in Section 4.1, in order to capture this aspect: (i) the dataset is organized by groups of museums having comparable popularity, where popularity is measured in terms of magnitude of number of followers of the account; (ii) for each group, the 20% of the tweets that were retweeted the highest number of times constitute the GOOD tweets, and analogously the 20% that were less retweeted are the BAD ones (the tests will also analyze the effect of changing the default class size of 20%).

3.2 Suggestions Mechanism

Whenever a draft message is classified as BAD, the media manager is presented some suggestions in order to enhance the message and, hopefully, make it more successful. To have suggestions easy to understand and fast to implement, we consider simple enhancements such as increasing or decreasing the value of the involved features (e.g., add/remove an image or a mention). Intuitively, our suggestion mechanism looks for GOOD tweets that are close to the BAD draft and that have been posted by museums with similar number of followers. Next, it seeks for differences in the features values and suggests to change the draft to make such values closer.

In details, given a draft message t, suggestions are produced following these steps:

[4]`http://darenr.github.io/afinn`

(1) *"Good" neighborhood search*: we determine the set N_t of GOOD neighbouring messages of t that are at a distance at most T from t. To this aim, each message is represented by the features selected for tweet classification. Feature values are computed by normalizing the corresponding raw values in the range [0,1]. Then, we use the Euclidean distance to compute distances among pairs of messages. The value of the threshold T is set as the average distance among all BAD tweets and their closest GOOD neighbour. In the unlucky case in which N_t is empty, we repeat the search by increasing the threshold to $3T/2$. If T_n results empty again, then the suggestion generation terminates and no suggestions are returned, since no useful suggestions could be given if not starting over again (possible good examples would be too distant from t and, thus, difficult to imitate without revolutionizing the message);
(2) *Definition of target features*: we analyze the classification of the draft message t to determine which features mostly concurred in the BAD classification, by means of their importance values. To this end, we implement the machine learning interpretability mechanism explained in Section 5. Such features are called *target* features as they are the ones that should be improved;
(3) *Computation of target feature values*: for each target feature, we define a target value by computing the mean of the raw values of that feature for the tweets in the neighborhood N_t, rounded to the closest integer. The mean is representative of the general trend of the "good" neighborhood;
(4) *Suggestion generation*: the target features values are compared to the features values of the message t and suggestions are provided to make the latter closer to the former. Suggestions are ordered by placing first suggestions concerning features that were the most determinant to the BAD classification.

An example of the suggestion generation process is given in Figure 1, in which t is the green circle. In step (1) three GOOD tweets nearest to t are identified within the given distance T In step (2) target features are selected according to the result of the classification of t. Figure 2 shows a graphical representation of the classifier output concerning those features that contributed the most to the classification of a tweet: dark pink segments refer to the features pushing toward being classified as GOOD, blue segments toward being classified as BAD. The length of each segment represents how much determinant the value of the corresponding feature had been for the final classification, while the value assigned to features is their current value for the classified tweet (e.g., the tweet contains 7 URL and zero mentions). In this example we can see that the number and the lengths of blue segments are larger than the number and lengths of the pink ones and, in effect, the tweet has been classified as BAD. Moreover, we can see that the feature that most influenced the classification as BAD was the number of URLs (being the corresponding blue segment much longer than any other segment), followed by the number of images. On the other side, sentiment, number of mentions and having a question mark were pushing for a classification as GOOD. Therefore, the selected target features are NURLS and NIMG. In step (3) target feature values are computed: we compute the mean of the NURLS values and of the NIMG values of the three GOOD tweets in the neighborhood, and we get that target value for NURLS is 4 and for NIMG is 2. Finally, step (4) generates the suggestions such as *"You should drop at least two or three URLs"* and *"You should add at least one image"*.

3.3 Dashboard Overview

We developed a prototype dashboard that implements our method. The dashboard is for mobile devices, so that it can be easily accessible to a large public of users using a widely available device.

The architecture consists of a *client side*, an Android application providing UI and communication with the server, and a *server side*, a server application accomplishing the *"intelligent"* tasks discussed in Section 3. From the user side, the dashboard design is meant to be intuitive, so that is can be accessible to a multiplicity of users with different technical skills (see Figures 3 and 4):

- at login (Figure 4-a), the user chooses his/her own museum, or selects a group of similarly sized museums in case his/her museum is not in the list, where similar size means similar number of followers of the museum

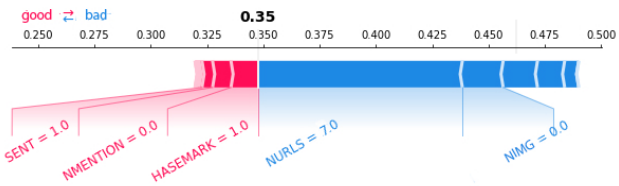


Fig. 2. Suggestion generation: graphical representation of the classifier output including features' contributions

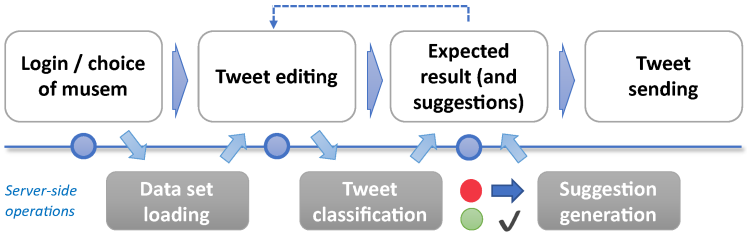


Fig. 3. Dashboard usage flow

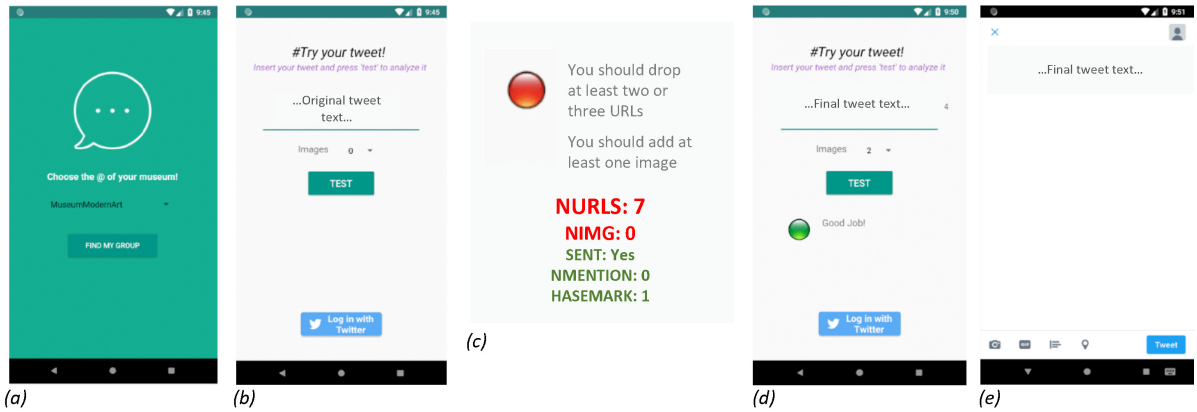


Fig. 4. Dashboard UI and usage example

Twitter accounts. The list of museums is dynamically filled at login on the basis of the data available in the server;

- the user edits his/her own draft message, and submits it for evaluation (Figure 4-b);
- the user gets a classification (either GOOD or BAD) for his/her draft message graphically shown by a green or red button, respectively;
- once the user is satisfied with the outcome (Figure 4-d), he/she can log onto Twitter and post the message (Figure 4-e).

Table 1. Museum groups with indication of number of followers (late spring 2020)

Group 1 (G1)	# Followers	**Group 2 (G2)**	# Followers	**Group 3 (G3)**	# Followers
@MuseumModernArt	5,384K	@britishmuseum	2,030K	@NationalGallery	884K
@Tate	4,874K	@saatchi_gallery	2,025K	@MuseeOrsay	707K
@metmuseum	4,364K	@vangoghmuseum	1,649K	@mfaboston	401K
@Guggenheim	3,502K	@MuseeLouvre	1,484K		
		@GettyMuseum	1,318K		
		@museodelprado	1,256K		
		@CentrePompidou	1,071K		
Group 4 (G4)	# Followers	**Group 5 (G5)**	# Followers		
@artinstitutechi	324K	@museupicasso	68K		
@museiincomune	279K	@mart_museum	65K		
@philamuseum	263K	@Uffizi	37K		
@ngadc	244K	@MuseoEgizio	28K		
		@MUSE_Trento	14K		

4 RESULTS

In this section we present the results of the experiments we performed to asses the validity of our method. Tests are performed on the prototype dashboard described in the previous section: the client side is implemented in Android studio and exploits the Volley library for networking. The server side is written in Python and exploits the Pandas library[5] for data manipulation, and the Scikit-learn[6], XGBoost [4] and SHAP libraries [18] for machine learning and classification functionalities. The dataset[7] and code[8] are publicly available.

4.1 Dataset

We selected 23 well known world spread art museums and grouped them into five groups: G1 (museums with at least three million of followers); G2 (museums with more than one million of followers); G3 (museums with more than 400,000 followers); G4 (museums with more that 200,000 followers); G5 (Italian museums). In this way, the analysis will not be influenced by the characteristics of a single museum. Table 1 reports the details of the grouping process. From these museums, we analyzed ca. 40,000 tweets, with a number varying from 5k ca. to 11k ca. for each museum group (see Table 2), depending on the number of museums in each group. Tables 3 and 4 report statistics on the value of context, content and output features for all tweets; see tables captions for details on the specific features and names. User features are MUSEUM, i.e., the name of the museum, and FOLLOWERS, the number of followers of the user posting the tweet. Moreover, given the rationale and the default class size of 20% discussed in Section 3.1, we have that a tweet is GOOD if it has more than 57, respectively 96, 30, 13 and 12 retweets for museum belonging group G1, respectively G2, G3, G4 and G5; a tweet is BAD if it has less than 7, respectively 9, 5, 2 and 1 retweets for museums belonging to group G1, respectively G2, G3, G4 and G5.

[5]http://pandas.pydata.org/
[6]http://scikit-learn.org/
[7]https://doi.org/10.5281/zenodo.4782984
[8]https://github.com/rmartoglia/predict-twitter-ch

Table 2. Number of tweets per museum group

	G1	G2	G3	G4	G5	Total
# Tweets	6097	11017	4936	7808	8233	38091

Table 3. Tweet Countable Content Features (number of hashtags, URLs, images, mentions and message length, respectively), Context Features (part of day and retweet) and Output Feature (number of retweets) (see Section 3.1 for description). `Min` is not reported as it is equal to zero for all groups and each feature, with the exception of length that varies from 7 to 19 characters. As for PARTOFDAY we have: morning (5:00- 11:59) = 0, afternoon (12:00 - 17:59) = 1, evening (18:00 - 22:59) = 2, night (23:00 - 4:59) = 3. ISRETWEET is 1 if the message is retweeted.

	Countable Content features					Context features		Output
	NHASH	NURLS	NIMG	NMENTIONS	LENGTH	PARTOFDAY	ISRETWEET	NRETWEET
	G1							
`Average`	0.96	0.80	2.47	0.50	225	0.97	0.11	68.35
`Std`	0.91	0.45	1.31	0.95	66	1.01	0.32	818.26
`Median`	1	1	2	0	238	1	0	23
`Max`	9	4	33	13	538	3	1	47558
	G2							
`Average`	1.34	0.60	2.56	0.49	230	0.94	0.24	75.42
`Std dev`	0.48	0.53	1.64	0.91	80	0.96	0.42	290.11
`Median`	1	1	2	0	239	1	0	28
`Max`	18	4	111	10	821	3	1	18092
	G3							
`Average`	1.04	0.68	2.51	0.55	246	0.83	0.15	21.9
`Std dev`	1.00	0.50	1.01	0.85	76	0.99	0.34	34.99
`Median`	1	1		0	267	1	0	13
`Max`	11	3	15	9	663	3	1	950
	G4							
`Average`	1.20	0.65	2.60	0.56	236	0.83	0.16	11.51
`Std dev`	0.48	0.51	1.32	1.16	71	0.97	0.36	112.45
`Median`	1	1	2	0	256	1	0	6
`Max`	15	3	17	16	643	3	1	9511
	G5							
`Average`	2.26	0.54	2.29	0.74	235	0.95	0.25	11.26
`Std dev`	2.05	0.56	1.20	1.29	81	0.94	0.43	139.08
`Median`	2	1	2	0	244	1	0	3
`Max`	18	4	38	16	579	3	1	11710

4.2 Tweet classification assessment

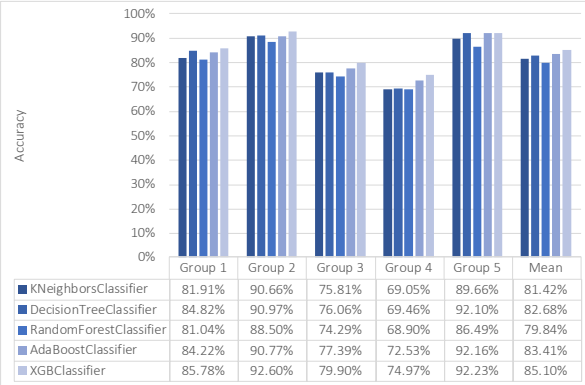
The machine learning approach for automatic message classification has been implemented by evaluating different ML models, including the most "popular" offered in the Scikit-learn and XGBoost libraries. We performed several experiments in order to evaluate not only the different classifiers, but also different sets of features and different class sizes and number. All non-numerical features (i.e., PARTOFDAY and MUSEUM) are encoded with standard

Table 4. Tweet On-Off Content Features (presence of exclamation mark, question mark, person names, place names, organization names, other names, topic density and sentiment) (see Section 3.1 for description). All values are binary. Minimum and maximum are not shown as they always are, respectively, 0 and 1. Density equal to 0 means the tweet is not dense, 1 it is. Sentiment equal to 0 means the tweet has no sentiment, 1 it has.

	On-Off Content features							
	HASEMARK	HASQMARK	HASNERPER	HASNERLOC	HASNERORG	HASNERMISC	DENSE	SENTIMENT
	G1							
Average	0.20	0.10	0.60	0.65	0.34	0.80	0.04	0.67
Std	0.42	0.32	0.49	0.47	0.48	0.40	0.19	0.47
Median	0	0	1	1	0	1	0	1
	G2							
Average	0.26	0.16	0.69	0.59	0.30	0.74	0.07	0.46
Std	0.43	0.34	0.47	0.48	0.45	0.43	0.25	0.50
Median	0	0	1	1	0	1	0	0
	G3							
Average	0.30	0.17	0.65	0.60	0.30	0.81	0.03	0.58
Std	0.45	0.39	0.48	0.48	0.46	0.38	0.18	0.50
Median	0	0	1	1	0	1	0	1
	G4							
Average	0.15	0.07	0.62	0.60	0.38	0.82	0.07	0.48
Std	0.35	0.26	0.49	0.47	0.48	0.38	0.26	0.50
Median	0	0	1	1	0	1	0	0
	G5							
Average	0.33	0.13	0.60	0.57	0.29	0.72	0.14	0.22
Std	0.46	0.35	0.49	0.49	0.45	0.45	0.35	0.41
Median	0	0	1	1	0	1	0	0

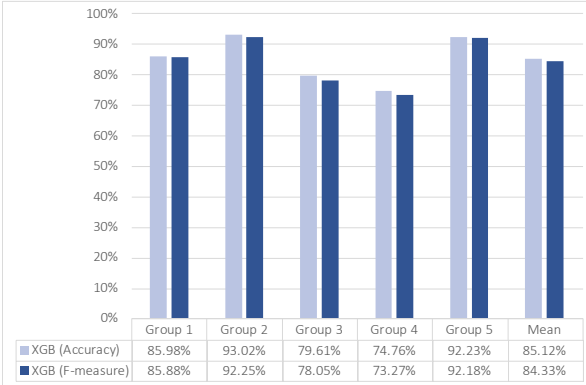
label encoding. In our default setting, the GOOD and BAD classes are defined with a threshold of 20% of the total messages as discussed in Section 3.1. Note that the dataset is naturally balanced between the classes due to its construction (equal number of GOOD and BAD samples). Training set and test set are chosen from the considered set of tweets in ratios of 9:10 and 1:10 respectively, w.r.t. the selected number of tweets. They are equally kept balanced by means of the stratified randomized folds cross validation technique, where the folds preserve an equal percentage of samples for each class. Classification performance is evaluated by means of accuracy figures, which is perfectly suitable for our balanced scenarios (anyway we will also briefly discuss F-measure performance). All the performance figures are derived as an average of 50 runs.

The first experiment is meant to select a proper classifier: we considered the complete features set (i.e., 17 features including 13 content features, 2 context features and 2 user features) and investigated the accuracy of the different classifiers: K-Nearest Neighbor (KNeighborsClassifier), Decision Tree (DecisionTreeClassifier), Random Forest (RandomForestClassifier), Adaptive Boosting (AdaBoostClassifier) and Gradient Boosting (XGBClassifier). Figure 5.a shows the obtained results: the accuracy (percentage of correct classifications) levels are quite high, especially for some of the groups/classifiers. In particular, results vary from group to group, but we can clearly see that the classifier offering overall better (and very convincing) results is the Gradient Boosting one: its accuracy ranges from 74.97% (Group 4) to 92.60% (Group 2), with a mean of 85.10% among all groups, and is consistently the



	Group 1	Group 2	Group 3	Group 4	Group 5	Mean
KNeighborsClassifier	81.91%	90.66%	75.81%	69.05%	89.66%	81.42%
DecisionTreeClassifier	84.82%	90.97%	76.06%	69.46%	92.10%	82.68%
RandomForestClassifier	81.04%	88.50%	74.29%	68.90%	86.49%	79.84%
AdaBoostClassifier	84.22%	90.77%	77.39%	72.53%	92.16%	83.41%
XGBClassifier	85.78%	92.60%	79.90%	74.97%	92.23%	85.10%

(a) Performances of different classifiers in terms of accuracy.



	Group 1	Group 2	Group 3	Group 4	Group 5	Mean
XGB (Accuracy)	85.98%	93.02%	79.61%	74.76%	92.23%	85.12%
XGB (F-measure)	85.88%	92.25%	78.05%	73.27%	92.18%	84.33%

(b) Performance of the Gradient Boosting classifier in terms of accuracy and F-measure.

Fig. 5. Classifier assessment

best one among all groups. For these reasons, Gradient Boosting has been chosen in the prototype implementation and the following tests will be performed by means of the this classifier alone. Figure 5.b also shows the F-measure metric results on the different groups: as we can see, these results are perfectly in line with the accuracy ones, due to the balanced data sets, therefore in the upcoming tests we will only show and discuss accuracy scores.

To verify the impact of features on accuracy values, we performed the second set of experiments employing different sets of features (i.e., content, context and user features). In particular, we removed context (i.e., PARTOFDAY and ISRETWEET) and user (i.e., FOLLOWERS and MUSEUM) information to understand the impact in the classifier prediction accuracy. Figure 6 shows the obtained results. It can be noted that the user features, as expected, help the classifier by better specifying the sender impact and therefore introduce important benefits (+11% for Group 1; +8% for Group 2; +6% for Group 3; +2% for Group 4; +7% for Group 5). The impact of context features appears, instead, more subtle (around +1% across all groups). Note that, even in the worst case where context and user information are not available/exploitable, the obtained accuracy with content-only features is still more than reasonable (ranging from 70.28% for Group 1 to 82.32% for Group 5). Consequently, we can state that, even if the content features typically bring the most important contribution in discriminating tweets, using user (i.e., source) information, when possible, can bring good accuracy improvements; the same applies to context information, even if to a lesser extent.

The third set of experiments investigates how accuracy varies w.r.t. different compositions of the considered dataset in class sizes and in different number of classes. In particular, we first evaluated the impact of considering larger parts of the original datasets by extending the GOOD and BAD classes beyond the 20% threshold. The obtained results are shown in Figure 7.a. As expected, gradually including a larger number of tweets that are more difficult to classify (i.e., closer to the "gray" zone) leads to a (slow) degradation of the performance. When considering the complete dataset (i.e., GOOD and BAD thresholds set to 50%), we loose about 15 points percentage on mean, even if the final accuracy can still be considered as quite satisfying (especially for some groups, e.g., near 80% for Group 2 and Group 5). As to the number of classes, we also wanted to measure the accuracy for a more complex (even if less frequent) scenario where it is needed to discriminate between a larger number of classes. More specifically, we considered the problem of classification among 3 classes, i.e., GOOD, BAD and a third AVERAGE ("gray" zone) class centered around the median of the number of retweets, with class sizes

ranging from 10% to 30%. Figure 7.b shows the obtained results. We can see that accuracy level, even if lower than for 2 class classification, is still convincing (above 64% on mean for the 20% class size), especially if we consider that, in this setting, a random classifier baseline would achieve 33% accuracy (rather than 50% accuracy as in the 2 class tests).

4.3 Message suggestion assessment

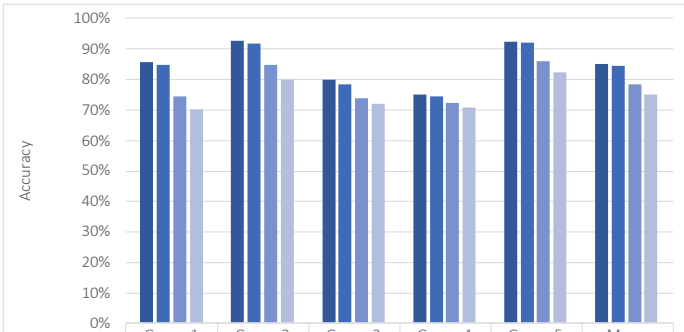
To assess the message suggestion mechanism, we set up a simulation experiment. Indeed, it is a hard task to test the mechanism in the real world as it would require to tweet from a museum account. Therefore, we simulated the behaviour of a media manager whose draft message has been classified as BAD. In our default setting, for each group, and for each BAD message t of the group test set:

(1) we compute suggestions for t;
(2) we apply the proposed suggestions (starting from the first, i.e. strongest, one, and stopping after a specified number of changes is performed to the message) and we get a new draft t';
(3) we classify t' and see if the classification has changed to GOOD.

Finally, we compute the percentage of t's that are classified as GOOD after applying the suggestions. Figure 8.a shows the obtained results for each group (and the mean between all groups) after a given number of suggestions/changes applied to the messages: as we can see, even when applying only the first proposed suggestion, nearly 60% of the BAD tweets are successfully classified as GOOD. When more than one suggestion is applied, the probability to produce successful tweets grows up to nearly 90% on average in the case of 4 applied changes. These promising results are also confirmed by the test shown in Figure 8.b, where we compared the effect of applying a given number of changes proposed by the suggestion mechanism to a random baseline where the same number of changes are applied in a random way (mean of 5 groups shown): in this latter case, even after 4 changes, only 14.92% (rather than 88.44%) of the messages have actually been improved to GOOD.

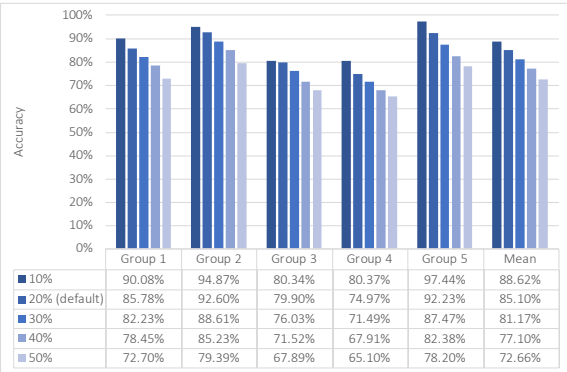
5 DISCUSSION: MODEL INTERPRETATION

As discussed in Section 4, our default (and best performing) configuration is based on the Gradient Boosting classifier offered by the XGBoost library. This algorithm, when employed by itself, works as a "black box" and is not able to deliver interpretability about the results (i.e., why a classification decision has been taken). In order to

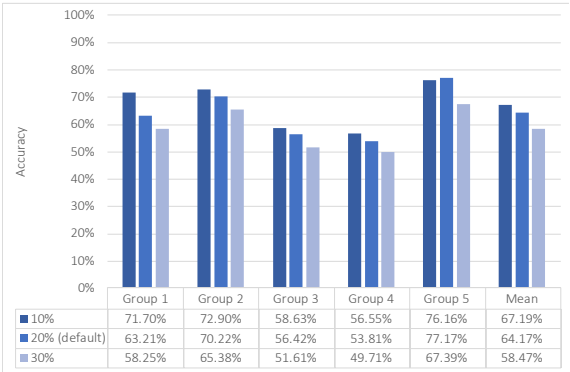


	Group 1	Group 2	Group 3	Group 4	Group 5	Mean
All (default)	85.78%	92.60%	79.90%	74.97%	92.23%	85.10%
No context	84.78%	91.89%	78.37%	74.43%	92.19%	84.33%
No user	74.34%	84.91%	73.69%	72.20%	85.96%	78.22%
Content only	70.28%	79.80%	72.07%	70.89%	82.32%	75.07%

Fig. 6. Accuracy comparison: the influence of context and user features vs content features.

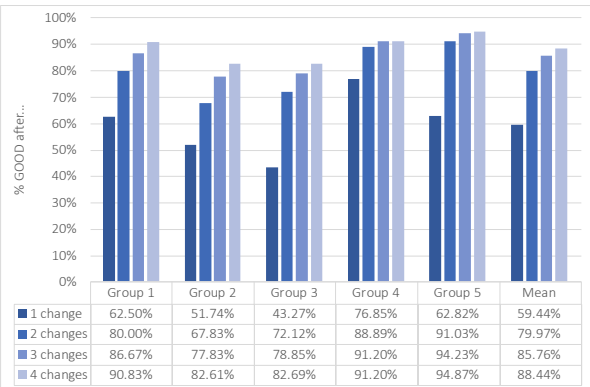


(a) 2 classes

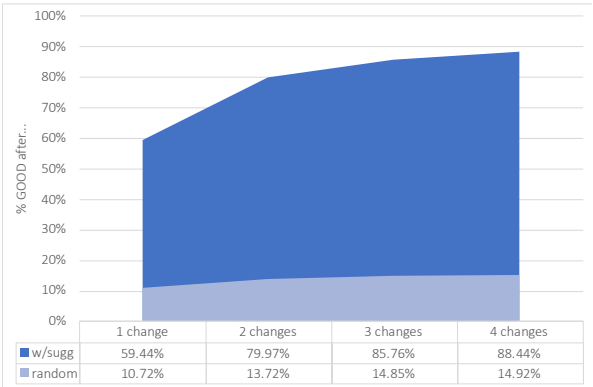


(b) 3 classes

Fig. 7. Accuracy comparison on different dataset compositions by varying the classes size.



(a) Impact of suggestion mechanism



(b) Comparison with random modifications (mean of 5 groups).

Fig. 8. Message suggestion impact

gain insight on the classification process and on the features that most influence it, we combined XGBoost with SHAP implementation [18], a library enabling the use of *SHapley Additive exPlanations (SHAP) Values* [23], first introduced in 1953 in the context of cooperative game theory. In this way we were able to generate reports on the relative importance of each individual feature, both across the entire dataset (global contribution) and for individual tweets (local contribution). The former are important to assess our classifier, the latter the suggestion mechanism.

Table 5 shows the feature importance (i.e., SHAP value) for all our features, for each group and in total as a sum of the scores of all five groups (most important features on top). Being interested in analyzing the global contribution of each feature to the classification of the tweets in each group, such global contribution of a feature passes through each individual sample of the training set. By computing the SHAP Values of each feature for each sample of the training set, the average per feature of the respective SHAP Values is calculated, thus generating a realistic summary of how much each feature has contributed to the classification. As we can see, feature

Feature	Group 1	Group 2	Group 3	Group 4	Group 5	TOT
MUSEUM	0.1558	0.1701	0.0045	0.0825	0.0580	0.47
FOLLOWERS	0.0575	0.0579	0.1342	0.0293	0.1252	0.40
NURLS	0.0050	0.0090	0.0079	0.0840	0.0988	0.20
ISRETWEET	0.0475	0.0205	0.0167	0.0017	0.0379	0.12
LENGTH	0.0168	0.0103	0.0325	0.0150	0.0163	0.09
NMENTION	0.0054	0.0459	0.0232	0.0075	0.0030	0.08
NIMG	0.0226	0.0047	0.0169	0.0231	0.0068	0.07
NHASH	0.0130	0.0222	0.0128	0.0093	0.0048	0.06
HASNERPER	0.0048	0.0096	0.0314	0.0016	0.0028	0.05
PARTOFDAY	0.0043	0.0085	0.0082	0.0171	0.0022	0.04
HASEMARK	0.0096	0.0003	0.0032	0.0030	0.0019	0.02
SENT	0.0035	0.0022	0.0007	0.0032	0.0059	0.02
HASQMARK	0.0003	0	0.0018	0.0006	0.0052	0.01
HASNERORG	0.0002	0	0.0018	0.0053	0.0001	0.01
HASNERLOC	0	0.0010	0.0038	0.0014	0.0003	0.01
HASNERMISC	0	0	0.0026	0.0016	0.0014	0.01
DENSE	0	0.0002	0.0023	0.0006	0	0.00

Table 5. Detailed impact of the different features (feature importance score): the higher the score the higher the impact.

importance varies from group to group, endorsing our choice to group museums. For example, the number of URLs is a factor at least 10 more important for classification in Groups 4 and 5 than for in others, and the number of mentions, NMENTIONS, is more than a factor 10 important for classification in Group2 than for Group 5. Anyway we can also see that there are features having a major role on the classification. The four most important features across groups are: the name of the museum (MUSEUM) and the number of followers (FOLLOWERS), the number of URLs (NURLS), the retweet flag (ISRETWEET). All in all, zero-importance features are almost absent, showing that most features give a contribution (even if, in case of the last ones, a very small one) to the classification in all groups.

Knowing the importance of the different features also enables the possibility of selecting only the top features and discarding the less relevant ones for each group. Figure 9 shows the accuracy comparison, in the default configuration, between the full feature set and a feature set only including the top 12 features. The accuracy is almost unchanged, confirming that: (a) the least important features give a very modest contribution; (b) feature selection does not improve our results, showing that in our scenario the Gradient Boosting classifier is very robust even to the inclusion of less informative features.

Finally, we proceed in interpreting the predictions of our machine learning model to an even higher level of detail. By plotting the SHAP Values of each feature for each sample of the training set, we produce a summary plot showing not only how much but also *in which way* each feature has contributed. Figures 10 and 11 show two examples (for Group 1 and Group 3, respectively) of such graphs; the others are not shown but their meaning will be summarized below. In these graphs, the MUSEUM feature is omitted (not significant for this analysis due to the large number of different values) and the other categorical feature (PARTOFDAY) is one-hot-encoded in order to show the contributions of the four possible values (morning, afternoon, evening and night). The key to reading these graphs lies in the distribution of the dots to the right or left of the vertical axis of 0. Isolating each feature from the others, each dot is a sample of the training set, its color (from blue to red) is determined by the value of that feature within its specific domain. The position is determined on the basis of the SHAP Value value of that feature for that sample. Those on the left of the 0 have a negative SHAP value (and thus a negative contribution to the classification) and vice versa. Besides the order of the features (most influencing features

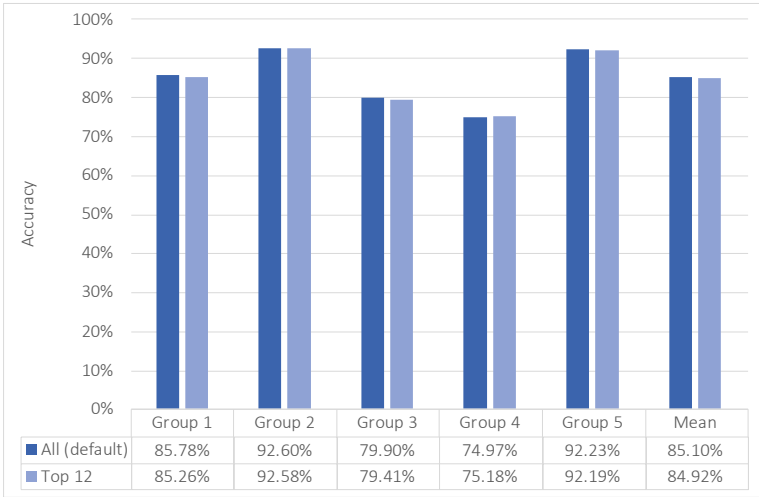


Fig. 9. Accuracy comparison with and without feature selection.

from top to bottom), which confirms what we already discussed in Table 5, the distribution and color of the dots allows us to identify in detail how the features are influencing the output.

Considering for example the graph for Group 1 (Figure 10), we observe that the LENGTH feature is clearly divided by the vertical axis: this indicates that all, or almost all, the samples that had reported a high value for the feature have a negative effect on the classification (and vice-versa). Even if for some features the dots distribution is not particularly explanatory (as in the case of the FOLLOWERS feature), by considering other features across the different groups we can draw some conclusions:

- the number of mentions (NMENTION) and URLs (NURLS) has a strong impact: a high number has generally negative effects on the classification (seen in all groups);
- the same holds for the number of images (NIMG) and hashtags (NHASH), particularly evident for Groups 1-3 and 5;
- tweets containing exclamation marks (HASEMARK) or DENSE tweets are generally not appreciated (especially in Groups 1 and 3);
- tweets citing famous people or, more generally, containing people names (HASNERPER) are more appreciated, while tweets citing specific locations (HASNERLOC) are less appreciated;
- the presence of sentiment (SENT) does not have a significant impact;
- as to context information, retwitted messages (ISRETWEET) have different impacts across groups (more appreciated in Groups 3-5, less in the others). Finally, the part of day (PARTOFDAY) seems to have an influence, in particular sending tweets during mornings and afternoons seems to have a positive impact on the classification (especially for Groups 1-4), while the opposite is true for evening and nights (evident for Group 1).

6 CONCLUSIONS AND FUTURE WORK

In this paper, we presented a machine learning method that can be used to predict the success of art museum tweets, where the prediction is based on content and context features; we also proposed a suggestion mechanism that might be used to enhance tweet drafts in order to increase their probability of success. By means of extensive experiments, we showed that the prediction is accurate and the suggestion mechanism effective. Indeed, it is

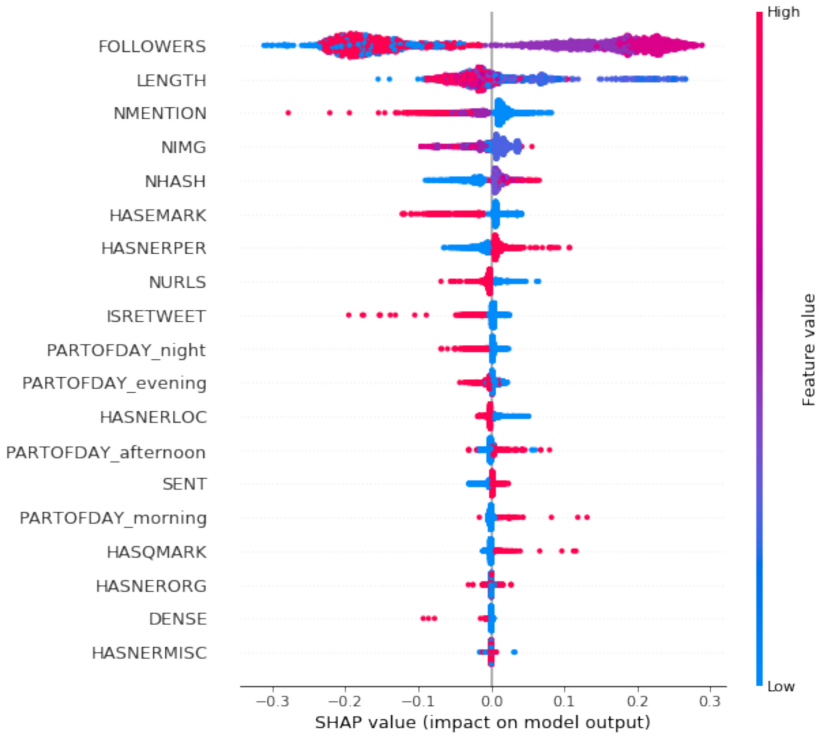


Fig. 10. Global interpretation of relevant feature impact for Group 1: each dot is a sample of the training set and is located the more to the left the more negative impact it has on the classification (or to the right for positive impact). The blue color represents low feature values (regardless if the impact is positive or negative), whereas red color represents high feature values.

not straightforward to find a correlation between tweets success and their content and context features; this is especially true when the analysis is to be performed at a mainly syntactic level, as was the focus of this paper, i.e., without examining the detailed semantic of the tweet sentences.

As for future work, we plan on investigating the possibility to include other features that might be derived from the message content and the importance of the linguistic aspects of tweets, in order to produce even higher accurate predictions. For example, tweet text messages might be analyzed in order to find the most representative words, or to understand if the message falls into some specific category (e.g., generic museum advertising, temporary exhibition announcement, invitation to special event, and so on). Finally, the above mentioned analyses will be exploited to empower the suggestion mechanism by making it provide actual content suggestions, e.g. to help writing the tweet with more effective words and sentence composition choices adequate to its category.

REFERENCES

[1] Muhammad Aurangzeb Ahmad, Carly Eckert, Ankur Teredesai, and Greg McKelvey. 2018. Interpretable Machine Learning in Healthcare. *IEEE Intelligent Informatics Bulletin* 19, 1 (2018), 1–7.

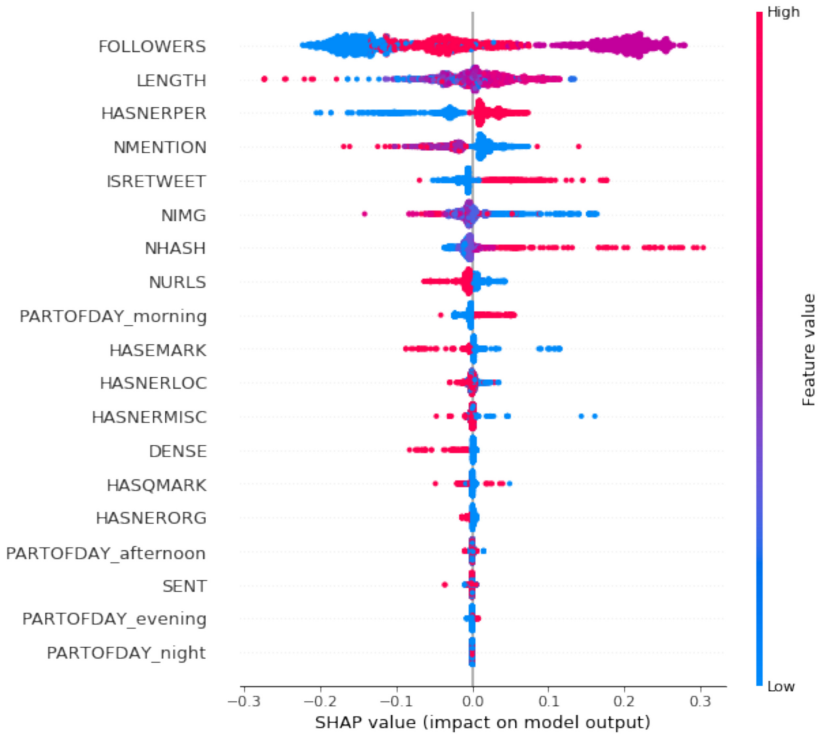


Fig. 11. Global interpretation of relevant feature impact for Group 3: each dot is a sample of the training set and is located the more to the left the more negative impact it has on the classification (or to the right for positive impact). The blue color represents low feature values (regardless if the impact is positive or negative), whereas red color represents high feature values.

[2] Stacy Baker. 2017. Identifying behaviors that generate positive interactions between science museums and people on Twitter. *Museum Management and Curatorship* 32, 2 (2017), 144–159. https://doi.org/10.1080/09647775.2016.1264882

[3] Roja Bandari, Sitaram Asur, and Bernardo A. Huberman. 2012. The Pulse of News in Social Media: Forecasting Popularity. In *Proceedings of the Sixth International Conference on Weblogs and Social Media.*

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proc. of ACM SIGKDD* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 785–794.

[5] Angelo Chianese, Fiammetta Marulli, and Francesco Piccialli. 2016. Cultural Heritage and Social Pulse: A Semantic Approach for CH Sensitivity Discovery in Social Media Data. In *Proc. of the IEEE 10th International Conference on Semantic Computing (ICSC)*. 459–464.

[6] Antoine Courtin, Brigitte Juanals, Jean-Luc Minel, and Mathilde de Saint Léger. 2014. The Museum Week Event: Analyzing Social Network Interactions in Cultural Fields. In *Proc. of the 10th International Conference on Signal-Image Technology and Internet-Based Systems*. 462–468.

[7] Davide Ferrari, Giovanni Guaraldi, Federica Mandreoli, Riccardo Martoglia, Jovana Milic, and Paolo Missier. 2020. Data-driven vs knowledge-driven inference of health outcomes in the ageing population: a case study. In *4th International workshop on Data Analytics solutions for Real-LIfe Applications, co-located with EDBT/ICDT 2020 Joint Conference (DARLI-AP @ EDBT 2020).*

[8] Adrienne Fletcher and Moon J. Lee. 2012. Current social media uses and evaluations in American museums. *Museum Management and Curatorship* 27, 5 (2012), 505–521. https://doi.org/10.1080/09647775.2012.738136

[9] Marco Furini, Federica Mandreoli, Riccardo Martoglia, and Manuela Montangero. 2017. The Use of Hashtags in the Promotion of Art Exhibitions. In *Digital Libraries and Archives - 13th Italian Research Conference on Digital Libraries (IRCDL), Revised Selected Papers.* 187–198.

[10] Marco Furini, Federica Mandreoli, Riccardo Martoglia, and Manuela Montangero. 2018. 5 Steps to Make Art Museums Tweet Influentially. In *Proc. of the 3rd International Workshop on Social Sensing (SocialSens)*.

[11] Marco Furini, Federica Mandreoli, Riccardo Martoglia, and Manuela Montangero. 2018. Towards Tweet Content Suggestions for Museum Media Managers. In *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good (Goodtechs '18)*. Association for Computing Machinery, 265–270.

[12] Shuai Gao, Jun Ma, and Zhumin Chen. 2014. Effective and Effortless Features for Popularity Prediction in Microblogging Network. In *Proceedings of the 23rd International Conference on World Wide Web*. 269–270.

[13] Anastasia Giachanou and Fabio Crestani. 2016. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Comput. Surv.* 49, 2 (2016), 28:1–28:41.

[14] Ion Gil-Fuentetaja and Maria Economou. 2019. Communicating Museum Collections Information Online: Analysis of the Philosophy of Communication Extending the Constructivist Approach. *J. Comput. Cult. Herit.* 12, 1, Article 3 (Feb. 2019), 16 pages. https://doi.org/10.1145/3283253

[15] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*. 80–89.

[16] Hiscox. 2018. Hiscox Online Art Trade Report 2019. *Hiscox Website. Available at: https://www.hiscox.co.uk/online-art-trade-report* (2018).

[17] Lesley Langa. 2014. Does Twitter Help Museums Engage with Visitors?. In *Proceedings of iConference*. 484–495.

[18] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774.

[19] Riccardo Martoglia and Manuela Montangero. 2020. An Intelligent Dashboard for Assisted Tweet Composition in the Cultural Heritage Area (Work-in-progress). In *Proc. of GOODTECHS20*. Association for Computing Machinery.

[20] Rebecca McMillen and Frances Alter. 2017. Social media, social inclusion, and museum disability access. *Museums & Social Issues* 12, 2 (2017), 115–125. https://doi.org/10.1080/15596893.2017.1361689

[21] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. In *Proc. of the 20th WWW*. 695–704.

[22] Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213 (2014), 1063–1064. https://doi.org/10.1126/science.346.6213.1063 arXiv:https://science.sciencemag.org/content/346/6213/1063.full.pdf

[23] Lloyd Stowell Shapley. 1953. *A Value for n-Person Games*. Contributions to the Theory of Games, Vol. 2. Princeton University Press, Chapter 17.

[24] Vasiliki Vrana, Kostas Zafiropoulos, and Konstantinos Antoniadis. 2016. Top European Museums on Twitter. In *Tourism and Culture in the Age of Innovation*. Springer, 457–469.

[25] Xiang Wang, Chen Wang, Zhaoyun Ding, Min Zhu, and Jiumin Huang. 2018. Predicting the popularity of topics based on user sentiment in microblogging websites. *J. Intell. Inf. Syst.* 51, 1 (2018), 97–114.

[26] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach. In *Proc. of the 20th ACM CIKM*. 1031–1040.

[27] Bo Wu and Haiying Shen. 2015. Analyzing and Predicting News Popularity on Twitter. *International Journal of Information Management* 35, 6 (2015), 702–711.

[28] Lun Zhang, Tai-Quan Peng, Ya-Peng Zhang, Xiao-Hong Wang, and Jonathan J. H. Zhu. 2014. Content or context: Which matters more in information processing on microblogging sites. *Computers in Human Behavior* 31 (2014), 242–249.

[29] Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. 2015. SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. 1513–1522.

[30] Xun Zhao, Feida Zhu, Weining Qian, and Aoying Zhou. 2012. Impact of Multimedia in Sina Weibo: Popularity and Life Span. In *Semantic Web and Web Science - 6th Chinese Semantic Web Symposium and 1st Chinese Web Science Conference*. 55–65.