IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 22, 2019, accepted November 13, 2019, date of publication November 15, 2019,
date of current version December 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2953856

Predicting Social Image Popularity Dynamics at Time Zero

ALESSANDRO ORTIS, (Member, IEEE),
GIOVANNI MARIA FARINELLA, (Senior Member, IEEE),
AND SEBASTIANO BATTIATO, (Senior Member, IEEE)
Department of Mathematics and Computer Science, Universit á degli Studi di Catania, 95125 Catania, Italy
Corresponding author: Alessandro Ortis (ortis@dmi.unict.it)

ABSTRACT This work addresses the task of forecasting the popularity achieved by images shared through social media over time. This task is known as ''Popularity Dynamic Prediction''. The work is motivated by the fact that the popularity of social images, which is usually estimated at a precise instant of the post lifecycle, could be affected by the period of the post (i.e., how old is the post). To this aim, we exploited a recently released dataset for popularity dynamic prediction that includes about 20.000 images uploaded on Flickr and their sequences of engagement scores (i.e., number of views, number of comments and number of favorites) with a daily granularity. To build such a dataset, each image and its accompanying meta-data and statistics are downloaded within a few hours from the image posting on the social platform. Then, an automatic procedure collected the daily engagement scores of each observed picture for 30 days. The paper presents an approach in which the engagement score is formulated as a composition of two information associated to the evolution over time (shape) and the order of magnitude (scale) of the sequence. The two properties are inferred individually, then the two results are combined to predict the popularity dynamics over n days. This paper presents exhaustive experiments on the addressed task, evaluating a large number of experimental settings for the predictions of popularity sequences with different length n ($n = 10$, 20 or 30). In all settings, the prediction performed by the proposed method can be computed before the image is posted (i.e., at time zero).

INDEX TERMS Dynamic prediction, image popularity prediction, social media engagement.

I. INTRODUCTION

In the context of social media analysis, there are several applications that could benefit from the assessment and the prediction of the level of engagement achieved by a post shared by a user on a social platform. Application examples are, among others, social media marketing, brand monitoring, and political parties popularity. The users engagement can be measured considering their activities and interactions with the content published on the social platform (e.g., comments, likes, views or shares). This information is often available and is usually compared with statistics of companies/advertisers websites and the users' queries on web search engines with the aim to assess the correlation between social advertising campaigns and their desired outcome (e.g., brand reputation, website/store visits, product dissemination and sale, etc.) [1]–[3]. The level of engagement of an image posted on a social network is usually referred as ''Image Popularity'' [4]. It is a difficult index to measure or predict. Several works aimed to find which features are correlated to image popularity within a social community or group of people with common interests [4]–[6]. Yamasaki *et al.* [7] proposed an algorithm to estimate the social popularity of images uploaded on Flickr by using tags. The method is used to recommend tags to users to gain greater attention from other users. The importance of each tag is obtained by combining tag frequency and weights. The results show that the popularity can be estimated from low dimensional (but meaningful) features such as image's tags. The exploited features are used for both regression (i.e., estimate the number of views, comments, favorites) and classification (i.e., distinguish between popular and unpopular images) tasks. In 2014, the authors of [4] proposed a popularity score for social media contents that is obtained by considering the cumulative engagement achieved up to the download time, normalized by the number of days since the content upload. This score is defined

The associate editor coordinating the review of this manuscript and approving it for publication was Hossein Rahmani.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

as following. Given an engagement score c_i achieved by an item i (e.g., number of comments, likes, views or shares) within T_i days since the content upload, the popularity score of i is computed as:

$$score(i) = \log\left(\frac{c_i}{T_i} + 1\right) \quad (1)$$

Khosla *et al.* [4] tried to understand the correlation between features extracted from the image or based on the user's statistics and the popularity score defined as in Equation (1). In particular, they considered the relevance of user context features (e.g., mean views, number of photos, number of contacts, number of groups, etc.), the image context features (e.g., title length, description length, number of tags), as well as visual features extracted from the image (e.g., GIST, LBP, CNN activations, etc.). Then, a Support Vector Regressor (SVR) has been trained exploiting the image and user features to estimate the popularity score. The performances have been measured in terms of Spearman's correlation. The authors of [8] formulated the task of image popularity prediction as a ranking problem. In particular, the authors used a latent SVM (Support Vector Machine) which objective function aimed to preserve the ranking of the popularity scores between pairs of images. They considered the number of views and the comments for Flickr images, whereas the number of re-tweets and favorites for Twitter posts. The problem of image popularity prediction was also addressed by Gelli *et al.* [9]. They used sentiment ANPs features (Adjective-Noun Pairs) defined in the Visual Sentiment Ontology (VSO) [10] pointing out that those features have a strong correlation with popularity. Aloufi *et al.* [11] evaluated several techniques to combine different features and investigated the effect of such combinations to predict different levels of interactions (i.e., number of views, number of comments and number of favorites on Flickr).

The work in [12] proposes a multi-modal approach that combines tags and visual features to predict the popularity (Eq. 1). In particular, the authors compared several unimodal and multi-modal settings in which features extracted from tags and Convolutional Neural Network (CNN) activations have been employed to train an SVR model. The results shown that tag features (i.e., sparse word count vectors) outperform visual features. In particular, the achieved results in terms of Spearman's correlation were lower than 0.3 when using visual features as unimodal input for SVR, lower than 0.5 when the concatenation of tag and visual features is used, whereas the best setting is obtained by considering only the tag feature (i.e., Spearman's correlation equal to 0.619). However, the visual features evaluated in [12] have been extracted considering different CNN architectures trained on the same image classification task [13].

The authors of [14] analysed the popularity of social posts published on Instagram by three different business accounts. Differently than common approaches that aims to estimate the popularity score defined in Equation (1), the authors of [14] aimed to predict a different score obtained by normalizing the number of views by the number of followers of the publisher account. This is motivated by the fact that in the specific context of business account analysis, the number of followers could augment significantly in a very short time. Beside the estimation of the popularity score, the authors of [14] also performed a classification on their data, by quantizing the popularity score into three categories: high, medium and low.

To the best of our knowledge, only few works considered the temporal evolution of the popularity (i.e., popularity dynamics). The work in [15] aims to predict the popularity stability of images published on Instagram. In particular, each image is classified as popular or unpopular, depending on both the number of likes and the age of the image. To this aim the authors defined different thresholds on the number of likes to categorize an image as popular or unpopular. The thresholds of popularity images have been defined considering the Pareto principle (80 % - 20 %) on the collected dataset. As instance, an image is considered popular if its number of likes is greater or equal than 49 for the first hour, 69 for the first day, or 75 for the first week. The features used in [15] include user's information, image semantics extracted from the captions of the images and the early popularity obtained in the first hour.

In [16] the authors considered the number of likes achieved after one hour from the image posting as input to predict the popularity score after one, seven or thirty days. The images have been classified as popular or non-popular. To this aim they used a popularity threshold obtained with the Pareto principle. Three features for the classification task were evaluated: social context (based on the user's number of followers), image semantics (based on image caption and NLP), and number of likes in the first hour. The binary classification has been performed by using a Naive Bayes Model with a Gaussian likelihood. The experimental results show that the number of likes of the first hour outperforms other features. Li *et al.* [17] extracted multiple time-scale features from a set of timestamps related to the photo post. The timestamp "postdate" is converted to several features with different time-scales: season of year, month of year, week of year, week of month, day of week, day of month, and moment of day, etc.

Wu *et al.* [18]–[20] explored social media popularity by modelling time-sensitive context and proposed to represent popularity in multi-time scales. The achieved results shown that the time of a post plays an important role for popularity prediction. A large-scale social media dataset, namely Social Media Prediction (SMP) dataset, has been collected to set the ACM Multimedia 2017 Social Media Prediction Challenge.[1] This dataset consists of over 850K posts and 80K users, including photos from VSO [10], as well as photos collected from personal users' albums. The first task (T1) of the SMP Challenge is the prediction of the popularity score as defined in [4], whereas in the second task (T2) a time-aware challenge is proposed. In particular, given the history of the past posts

[1]Challenge webpage: https://social-media-prediction.github.io/MM17PredictionChallenge

of a group of users of a social platform, the task T2 requires to predict the *top k* popular posts (i.e., the ranking) of a set of new posts. Differently than the T2 of the SMP Challenge, the dataset presented in [21] reports the engagement statistics of all the crawled images monitored daily for 30 consecutive days since their upload. In this way, is possible to compute the popularity score for an arbitrary day and try the forecasting of the overall sequence. Moreover, the sequence of daily scores can be exploited to analyse the post lifecycle. Previous works on time-aware popularity prediction employ temporal information (e.g., engagement in the first period) to infer the image popularity score as in Equation (1) at a specific time or at pre-defined time scales [16], [17]. In other words, these systems include the time information in the input, in order to better predict the popularity score defined as a single value.

In the context of modelling and predicting dynamics, several works have been presented. The authors of [22] aimed to predict the popularity evolution of user generated videos. They defined a set of popularity ''behaviours'' of the contents (i.e., patterns). The proposed method compares the popularity achieved by different contents up to the current time step, performing a clustering over the set of defined behaviour patterns. The clustering procedure is performed over time and, for each content, the system is able to predict the next clustering. In other words, given a set of contents (i.e., videos) and the clustering at time t, the system aims to predict the most likely cluster that each element will belong to at time $t+k$. The paper in [23] proposed two models to predict the future popularity of YouTube videos, by exploiting a set of daily samples of the content's popularity measures up to a given reference date, which are properly weighted. The work in [24] analyses the early patterns of YouTube videos and Digg stories to predict the long-term popularity of such contents. For instance, the information related to the users access to a a Digg content (i.e., number of votes) allow the system to the method to predict the popularity 30 days ahead with an error of 10%. To achieve the same error rate for the popularity prediction of YouTube contents, the system needs to know the information related to the first 10 days. The authors of [25] proposed a cross-domain approach that correlates popular topics on Twitter with YouTube videos on the same topics and compares the popularity of the same video on the two platforms to predict a sudden burst of popularity. The paper in [26] tries to estimate the popularity evolution of on-line videos. It models the popularity evolution by a function which parameters are estimated at video's early age by fitting its early view count trace. In [27] the number of views of a video is predicted via a regression approach. The method aims to predict the number of views at time t_i, given th features from the first t_j days after publication ($t_j < t_i$), by means of a Support Vector Regressor with Gaussian Radial Basis Functions. The paper in [28] presents a method to predict long-term popularity of User Generated Content by modelling the patterns of popularity dynamics. The popularity sequences are modelled by means of a Gaussian Mixture Model, where each component of the mixture represents a segment of the whole path. The method tries to predict each segment taking into account the past predictions and a number of known popularity values sampled from the sequence (e.g., the number of views in the first week). Indeed, the authors used the cumulative number of views (i.e., popularity) over time as only feature.

All the mentioned works aim to predict popularity by exploiting information related to the popularity achieved just in the first period (e.g., [15], [16], [28]), often referred as early popularity or early reaction. In [29] the authors addressed the problem of predicting the popularity of social posts before their upload. However, the experiments focused on news contents (i.e., textual tweets). Moreover, as claimed by the authors themselves, the results on popularity prediction were not satisfactory. In fact, the best results have been obtained by quantizing the values of popularity scores and training a classifier.

Differently than all previous works, our approach focuses on the prediction of the whole temporal sequence of image popularity scores (with length 10, 20 or 30 days) with a daily granularity, before any post's statistic (i.e., early popularity) is available. The challenge of predicting the popularity dynamics of social images has been introduced in [30], which presents an overview on image sentiment analysis and related tasks. We investigate the correlation between social image popularity, social features (including user's and photo's statistics), and visual features extracted from images. The set of features have been selected taking into account the insights achieved by the state of the art represented by the works that aim to predict the normalized popularity score defined in Equation (1). This paper extends our previous work [21], in which the popularity dynamics dataset has been presented, as well as a set of experiments aimed to explore the new proposed challenge. With respect to our previous work, this paper extends the problem analysis as well as the evaluation of the inference task by considering the prediction of sequences with length of 10 and 20 days, which revealed more challenging and interesting scenarios, as the popularity dynamic presents high variabilities in the first period. Furthermore, the experimental settings have been extended by considering several combinations and fusion strategies of the input features. The main contributions of this work are the following:

- it poses new questions around on-line behaviour, popularity, and social media content lifecycle;
- it addresses a very challenging task which finds several practical uses in the context of social media analysis applications such as recommender systems and advertisement campaigns analysis/placement;
- the proposed system allows the definition of applications to support the publication and effective diffusion of contents through social media, by implementing a forecast of the engagement evolution over time. As instance,

such a system can indicate when old contents should be replaced by new ones before they become obsolete.
- an in-deep investigation of social features that characterize the influence of a user and the level of diffusion of a photo is provided. In addition to the main user and photo features, we considered statistics related to groups in which the user is enrolled or in which the photo is shared;
- an ablation study of the problem aimed to understand how the prediction is affected by different information. Each case of study defined during the analysis is based on the knowledge of a part of the information to be inferred (i.e., Ground Truth), this allows to assess the contribute of each learned module of the proposed approach.

The paper is organized as in the following. We provide a study of the time effect on the problem of popularity prediction in Section II. Section III provides a description of the popularity dynamics dataset, by detailing the crawling procedure and the analysis of the crawled data, as well as the data preprocessing. Section IV gives the details of the proposed approach. The evaluation of the proposed method is reported in Section VI, whereas the obtained results are commented in Section VII, which also provides insights and cues for future works.

II. MOTIVATIONS

In Equation (1) the total engagement achieved by an image post is normalized by the age of the post (i.e., the number of days from its upload). However, such definition of popularity does not take into account the evolution of the image's level of engagement over time. It is possible to observe that the daily increment of the engagement scores over time is very high in the first period and tends to decrease with the passing of the days. As consequence, Equation (1) penalizes old contents with respect to more recent ones. Most of the engagements (e.g., views) obtained by a social media item are achieved in the first period. The study presented in [31] shows that photos shared on Flickr gain the majority of their engagements within the first week (early period). Figure 1 shows the dynamics of the number of views related to three different photos shared on Flickr. The number of views of each photo is extracted daily for a period of 30 days. We can observe that all three images' dynamics have similar shapes (especially the first and the second plot from the top), but they have different scales (i.e., the number of views after 30 days is different of about an order of magnitude). As observed by previous studies on the field of social media interactions [31], the number of views achieved by a social image usually increases in the first week, then is quite stable over time. This means that the popularity score introduced in [4] (i.e., Equation (1)) decreases with the age of the post. However this fact has not been considered in the popularity score formulation. As consequence, two image posts with similar engagement dynamic but very different evolution, are ranked differently depending on the time of analysis (i.e., download time). This is clear by considering Figure 2, which shows the popularity score of Equation (1) computed over time for the three examples depicted in Figure 1. Furthermore, the presented study also reveals that several engagement dynamic trends (i.e., engagement shapes) exist. This also affects the comparison of two picture by means of their popularity scores computed at different time instants.

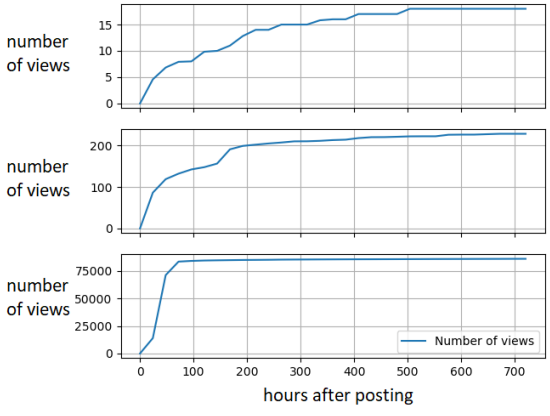

FIGURE 1. Examples of dynamics of the overall number of views achieved by three photos shared on Flickr related to 30 days. The shape of the three sequences is very similar, however they have very different scales.

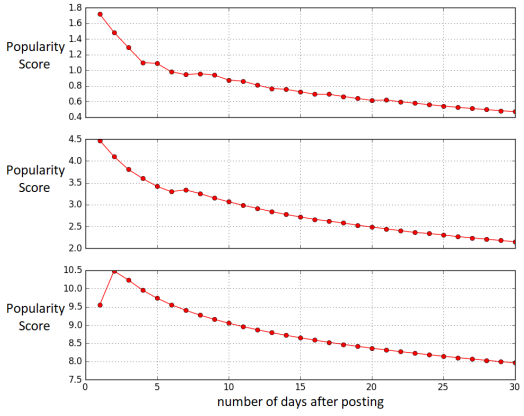

FIGURE 2. Popularity scores over time computed using Equation (1). After the first period the popularity score decreases with the post age. This effect is caused by the very low engagement around the photo after the early period.

For the above reasons, we addressed the problem of image popularity prediction taking into account the temporal dynamic of the engagement score (i.e., the variation of the engagement value over time).

III. POPULARITY DYNAMICS DATASET

The dataset has been obtained by exploiting the Flickr API,[2] which allows to retrieve images and the related information shared by users which are publicly available on the social platform. By means of a crawling process, more than 20.000 Flickr images have been downloaded and monitored for 30 consecutive days after their upload [21]. In particular, first the crawling process downloads a group of the latest

[2]https://www.flickr.com/services/api/

images published on Flickr, as well as a set of features related to the user's and photo's statistics. Then, the engagement scores (i.e., number of views, number of comments, and number of favorites) are downloaded daily for the following 30 days. In the experiments described in the following sections, the engagement score is computed considering the number of views. However, the daily number of comments and number of favorites are included in the dataset for extended studies.

Posts with a dimension of the image less than 10 Kb have been removed, since it usually corresponds to icons or place-holder images used by Flickr to indicate that the picture is no longer available.

The engagement score sequences of 30 days have been pre-processed in order to obtain values with regular time intervals of 24 hours. To this aim, we approximated the score function between two consecutive samples with a linear function. Figure 1 shows three examples of the score obtained with this process.

During the crawling process, a large group of statistics related to users, photos and groups have been collected. In particular, for each user, the following information have been collected: number of contacts, if the user is a professional photographer, number of photos, mean number of views, number of groups, the average number of people and images of the user's groups. The information related to the photo are: the original resolution, title length, description length, number of albums, number of groups, the average number of people and photos in the groups in which the picture has been shared in, and the social tags associated to the photo. For each social entity (i.e., users, photo and group) the IDs on the Flickr platform is also included in the dataset. In addition, for each picture, the GPS coordinates (when available) as well as the timestamps related to upload, download and acquisition date and time are included in the dataset. These information can be exploited to extend the crawling with more specific data and analysis. The collected dataset is publicly available.[3]

A total of 21.035 pictures have been crawled and monitored for 30 days. During the 30 days monitoring, some photos have been removed by authors (or not longer publicly available). As consequence, a lower number of photos have been tracked for 30 consecutive days. In particular the final dataset consists of:

- a total of 19.213 photos monitored at least for 10 days;
- a total of 18.838 photos monitored at least for 20 days;
- a total of 17.832 photos monitored at least for 30 days;

In the experiments described in the following sections, we first considered the set of photos monitored for 30 days. Then we extended the experiments on the other two sets. The average number of images per user in the dataset is lower than two (~ 1.6). Therefore, the dataset represents a setting where different images belong to different users. This is often the case in search engine results. Indeed, search engines can exploit popularity prediction systems in order to better rank the retrieved results.

[3]http://iplab.dmi.unict.it/popularitydataset/

IV. PROPOSED METHOD

Given an engagement sequence s of the number of views achieved by a Flickr photo over n days, the proposed framework models splits s into two properties: the *sequence shape* and the *sequence scale*. Specifically, these properties are defined in Equation (2). In particular, s_{scale} is defined as the maximum value of s, whereas s_{shape} is obtained by dividing each value of s by s_{scale}:

$$\begin{aligned} s &= [v_0, v_1, \ldots, v_n] \\ s_{scale} &= max\{s\} = v_n \\ s_{shape} &= \left[\frac{v_0}{v_n}, \frac{v_1}{v_n}, \ldots, \frac{v_n}{v_n}\right] \end{aligned} \tag{2}$$

Therefore, given s_{shape} and s_{scale}, is possible to obtain the original sequence s as follows.

$$s = s_{shape} \times s_{scale} \tag{3}$$

Note that, since each sequence represents a cumulative function, the last value of the sequence (i.e., v_n) always corresponds to the maximum value of the sequence s. Hence, the engagement sequence can be viewed as a pair of the two properties, the scale s_{scale} and the shape s_{shape}. The scale property is the degree of popularity achieved by the photo, considering the cumulative engagement after n days. The shape generalizes the temporal evolution of the sequence (i.e., its trend) in the observation period, regardless its actual values. In the proposed method, two separate estimations for the shape and the scale of the engagement have been performed. Then, Equation (3) is exploited to combine the two information and estimate the original engagement sequence associated to the photo.

The proposed method assumes the independent relationship between the scale and the shape of a sequence. Indeed, we observed that two sequences with the same shape can have very different scales (e.g., as the three examples depicted in Figure 1 of the paper), and vice-versa. To motivate this assumption, we analysed the distributions of the s_{scale} values related to sequences grouped by the assigned shape prototype (i.e., s^*_{shape}). Figure 4 shows the distributions of the s_{scale} values related to sequences grouped by the assigned shape prototype. (i.e., s^*_{shape}). There is a huge variability of the s_{scale} values within the sequences of the same shape. Note that the s_{scale} axis is represented in logarithmic scale, to better visualize the range of the distributions. This motivates our assumption that the two properties are unrelated.

Figure 3 shows the overall scheme of the proposed approach. First, a clustering analysis has been performed on the shapes of the training sequences. The result is a set of general shapes that we can exploit as prototypes. In particular, the centroid of each cluster, denoted as $s^*{}_{shape}$, is used as prototype for all the sequences in the same cluster. The obtained shape clusters are used as classes to define a classifier trained to predict the cluster to which a new sequence has

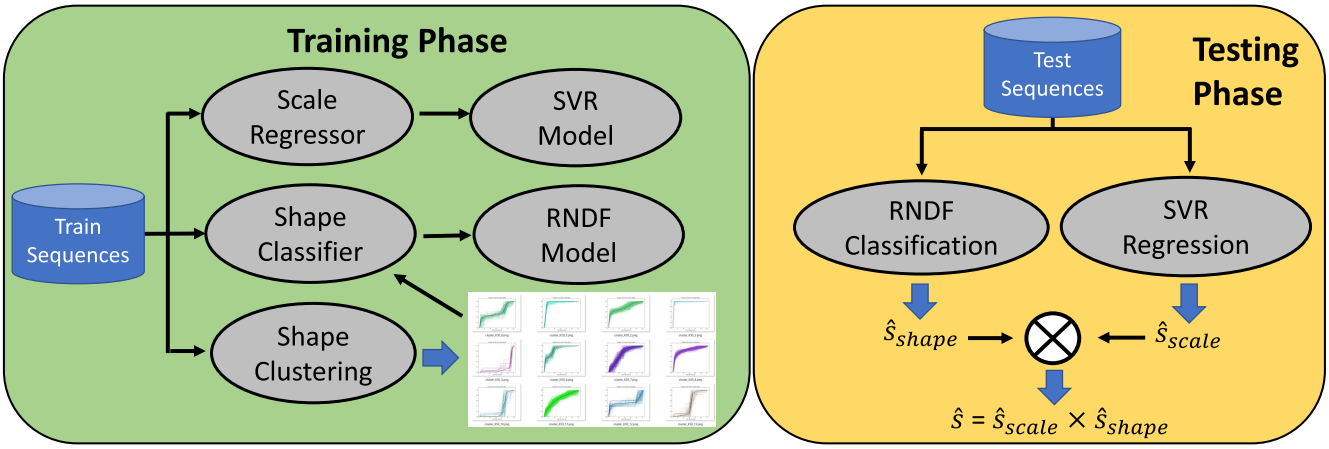

FIGURE 3. The proposed approach models the problem as the combination of two properties: the sequence engagement shape $\hat{s}_{shape}$ and scale $\hat{s}_{scale}$. The shape property is estimated by a Random Forest (RNDF) classifier, whereas a Support Vector Regressor (SVR) is employed to produce the scale factor $\hat{s}_{scale}$. The labels used to train the shape classifier are generated by mapping the shapes s_{shape} (Eq. 2) in one of the 50 shape prototypes obtained by clustering. This pre-processing step is performed after a clustering procedure over the training dataset (green backgroud). The regressor is trained by considering the value s_{scale} of the training set as Ground Truth.

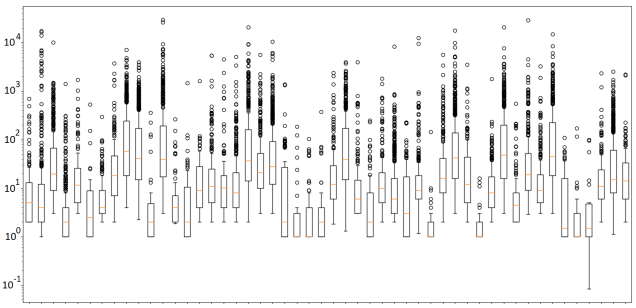

FIGURE 4. Distributions of the s_{scale} values of sequences grouped by the s^*_{shape} prototypes assigned by the clustering procedure.

to be assigned, given the set input social features. Since each cluster is associated to a shape prototype, a prediction corresponds to the selection of a shape prototype $s^*{}_{shape}$, among the set of prototypes defined by the clustering analysis. The predicted prototype for a new sequence is denoted by $\hat{s}_{shape}$. The sequence scale s_{scale} is obtained by means of a Support Vector Regressor (SVR), trained to estimate the value of the scale given a set of social features. The output of the SVR is denoted by $\hat{s}_{scale}$. The estimation of the engagement sequence of an image post for the period of n days is then obtained as $\hat{s} = \hat{s}_{shape} \times \hat{s}_{scale}$. To measure the performance of the system we use a test dataset and the Root Mean Squared Error (RMSE) between the Ground Truth sequences s of the test sequences and the predicted ones $\hat{s}$.

The experimental results reported in this paper have been obtained by considering 10 random train/test splits with 90% of images used for training and 10% used for test. All experiments, have been performed considering features based on the user statistics, photo information and visual content (i.e., extracted from the picture). Although previous works on popularity prediction shown that proper combinations of user statistics and photo information obtained the best results, they also suggest that the semantic content of the picture (i.e., presence of specific objects, scenes, etc.) may have an impact on the popularity prediction [4]. Considering the above observations, we additionally evaluated 6 visual representations extracted from the images by exploiting three state-of-the-art CNNs. For each model, we considered the last two activation layers, referred here as $f1$ and $f2$, before the softmax classifier. Specifically we considered the following architectures:

- **Hybridnet [32]**: specialized to classify images into one of 1.183 categories (978 objects and 205 places).
- **DeepSentiBank [33]**: specialized to assign an Adjective-Noun Pair to an input image among 4.342 different ANPs [10].
- **GoogleNet [34]**: specialized to perform image classification among 1.000 object categories.

Differently than previous approaches that exploited CNN activations as features for the popularity prediction task (e.g., [12]), we evaluated the activations of three CNNs trained on different tasks with different datasets. In the following we describe the details of each involved step.

A. SHAPE PROTOTYPING

Due to the formulation described in Equation (2), all the values of the s_{shape} sequences are in [0, 1]. We consider that all the sequences with the ''same'' dynamics will have a very similar s_{shape}. As groups of sequences with the same shape are examples of dynamics with a common engagement evolution, we fist tried to define a number of engagement prototypes, representing the different groups of shapes. In particular, we grouped the training normalized sequences (i.e., s_{shape}) by a K-means clustering. The resulting centroids represent the dynamic models for the sequences within each cluster (i.e., each shape prototype corresponds to a cluster centroid).

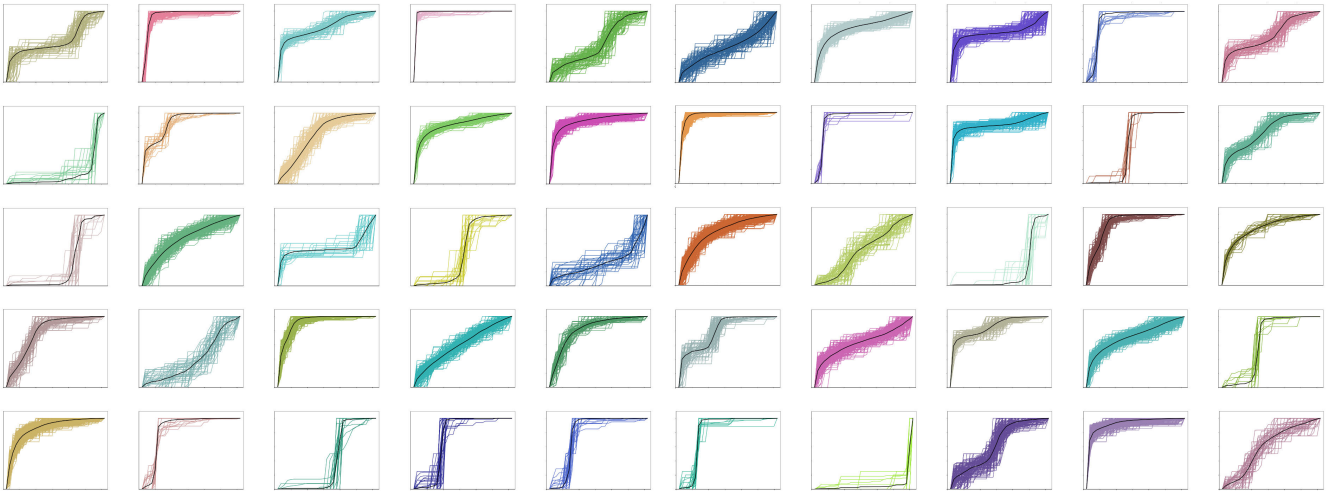

FIGURE 5. The considered shape prototypes obtained by clustering the training sequences (best seen in color).

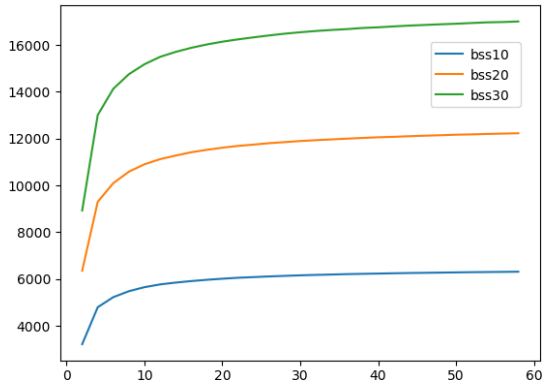

FIGURE 6. Between cluster Sum of Squares (BSS) analysis of the sequences with length 10 (blue), 20 (orange) or 30 (green) days.

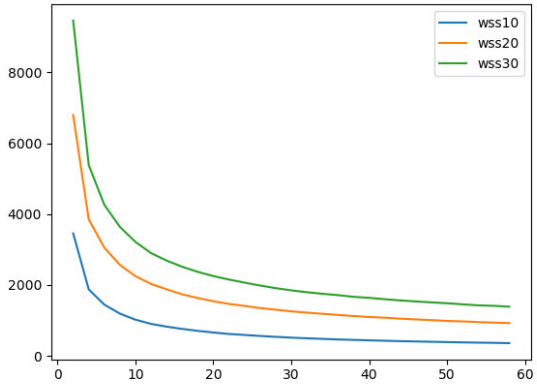

FIGURE 7. Within cluster Sum of Squares (WSS) analysis of the sequences with length 10 (blue), 20 (orange) or 30 (green) days.

In order to determine the best value for K, we performed the K-means algorithm evaluating a large range of values for K. Then, the best K has been selected by considering the Within cluster Sum of Squares (WSS) and the Between cluster Sum of Squares (BSS) scores obtained by varying the parameter K, which provide a measure of cluster cohesion and cluster separation respectively. In particular, we want the WSS value to be low and the BSS value to be high. The WSS/BSS analysis has been performed considering the groups of sequences of 10, 20 and 30 days. Figure 6 and Figure 7 show the values of BSS and WSS respectively. When K increases the cluster elements are closer to the cluster centroid. The improvements will decline, and at some point become more stable. We then tested a few values of K which corresponds to the plateau of the BSS/WSS functions.

Indeed, the experiments suggest more than one value for the optimal number of clusters. The best results have been obtained with $K = 45$ (10 and 20 days sequences) and $K = 50$ (30 days sequences).

Figure 5 shows the clustering result for the sequences of 30 days, with a set of plots. In particular, for each plot, the black bold line depicts the cluster centroid (i.e., the shape prototype), whereas the coloured lines represent the shapes of the training sequences belonging to the cluster.

B. SHAPE PREDICTION

After the clustering analysis described in Section IV-A we have a set of shape prototypes. This set can be considered a "dictionary" of the temporal dynamics for the popularity sequences. Indeed, any sequence can be assigned to a cluster by comparing its shape with respect to the prototypes. Each sequence of the training and test sets has been assigned to a shape prototype. Then, the training set has been used to train a classifier that considers only the features extracted from a social post, and predicts the shape of the corresponding sequence (i.e., the assigned prototype). A pool of classifiers has been evaluated to find the best classification algorithm. Moreover, during the evaluation we further considered several combinations of features to be used as

input for the classifier. In particular, The following classifiers have been evaluated: Decision Tree Classifier (DT), Random Forest Classifier (RNDF), SVM with RBF (RBFSVM) or linear (LSVM) kernel, k-Nearest Neighbour (kNN) and Multi-layer Perceptron Classifier (MLP). The best classification performances have been obtained by using all the social features as input combined with a RNDF Classifier. Hence, in the proposed approach, given a new test photo and its social features, a RNDF classifier is used to assign a shape prototype. The DT classifier obtains slightly lower results, whereas kNN and SVM achieved lower performances. The MLP Classifier resulted the worst approach, probably due to the limited number of examples for each class and the unbalanced distribution of the elements in the clusters. To deal with this issue we performed a stratified approach by considering 10 random splits in all our experiments. Considering stratification we ensure that each fold contains roughly the same proportion of the classes.

Most of the evaluated classifiers require the choice of specific parameters (e.g., the parameters C and γ for LSVM and RBFSVM, the number of neighbours K for kNN etc.). During our evaluation, the parameters have been determined by performing a grid-search method on the training data.

C. SCALE ESTIMATION

The estimation of the sequence scale s_{scale} is performed by a Support Vector Regressor (SVR) trained on the training sequences. In particular, in the proposed approach, the SVR has been trained on the normalized popularity scores defined as in [4]. Then, the following Equation (4) is exploited to transform the predicted popularity scores to the estimated number of views. Let $\hat{p}$ be the popularity score (i.e., Equation (1)) estimated by the SVR, the number of views is computed by applying the following formula:

$$\hat{s}_{scale} = (e^{\hat{p}} - 1) \times n \tag{4}$$

In our experiments, we considered several combinations of social features. First, we performed a single feature evaluation by training a SVR considering each feature individually. Each experimental setting has been evaluated by computing the Spearman's correlation between the predicted popularity score and the Ground Truth popularity (i.e., Equation (1)). This approach represents the common way to evaluate the classic popularity prediction approaches [4], [11] and to provide a correlation estimate between the features and the value of s_{scale}.

In the second stage of experiments we evaluated several combinations and compositions of the selected features. In particular, we trained a SVR by considering different concatenations of social features as input. Based on the performances obtained in the previous single feature evaluation, we defined proper groups of features. Furthermore, four approaches that takes the outputs of SVRs trained with the single features as input have been evaluated. Since these approaches perform a fusion of the outputs after the prediction, they are commonly referred as ''late fusion strategies''. In particular we evaluated the following approaches:

- **Late Fusion 1:** this approach computes the outputs of the considered SVRs;
- **Late Fusion 2:** this approach concatenates the outputs of the considered SVRs and trains a new SVR;
- **Borda Count:** the output is obtained by computing a weighted average of the single feature SVR outputs. The m evaluated features are ranked in descending order based on the achieved Spearman's correlation. The weight of the output corresponding to the top ranked feature is set to m, the second ranked feature is weighted with $m - 1$, and so on;
- **Weighted Fusion k:** the features are ranked in descending order considering the achieved Spearman's correlation. Then, the final output is obtained by computing a weighted average of the first k single feature SVR outputs. In this case the weights corresponds to the Spearman's correlations achieved individually.

The performances on scale estimation in terms of Spearman's correlation are reported in Table 2 and 3 (see Spearman column).

V. PROBLEM ANALYSIS

As previously observed, most of the state of the art approaches on image popularity prediction address a simplified task, defined by quantizing the range of output values into two categories [7], [16], [29], or by transforming the problem to a ranking task between pictures [8] instead of predicting the actual values of popularity. Furthermore, the majority of these works predict a single score, normalized at an arbitrary time, rather than the temporal evolution of the photo popularity. Moreover, there are just a few works that aim to predict dynamics. However, these methods require the knowledge of the first values of the sequence (e.g., [16]). This indicates that the popularity dynamic prediction is a very challenging task. In order to better comprehend the difficulty of the task, we performed an ablation analysis aimed to assess how the shape and the scale inferences affect the estimation of the popularity dynamic sequence.

We defined three different experimental settings. Each setting exploits a certain amount of knowledge about either the scale and the shape of the sequences (i.e., the Ground Truth scale and/or shape prototype). Therefore, the achieved error rates resulting from these experiments can be considered as lower bounds for any approach that aims to predict the popularity dynamic without any prior knowledge on the output sequence. The evaluated experimental settings have been defined as follows:

- **Case A:** in this case, the output sequence is obtained by combining the shape prototype assigned by the clustering procedure (s^*_{shape}) and the Ground Truth scale value (s_{scale}). Since both the values are taken from the

Ground Truth, this method achieves the minimum possible error. In particular, the error values measured under this setting are caused by the clustering approximation of the sequences.

- **Case B:** in this setting, the scale value is taken from the Ground Truth (s_{scale}), whereas the sequence shape (i.e., the assigned prototype) is predicted ($\hat{s}_{shape}$) by exploiting the Random Forest classifier (RNDF). The measured error caused by the clustering approximation of the sequences and the error of the Random Forest classifier that tries to predict the assigned shape prototype.
- **Case C:** this case exploits the shape prototype assigned as Ground Truth (s^*_{shape}) and combines it with the scale value inferred by the trained Support Vector Regressor ($\hat{s}_{scale}$). In this case, the error is caused by the clustering approximation of the sequences and the estimation error of the SVR in the inference of the scale value.

The mean RMSE errors of Case A and Case B are 4.97 and 7.51 respectively. Since Case A and Case B exploit the Ground Truth scale values, the results achieved on these two experimental settings are not affected by the scale inference and depends only on the clustering and shape prototyping steps (see Table 1). The error rates of Case C, instead, are affected by the estimation of the sequence scale.

TABLE 1. RMSE errors of Case A and Case B experiments. These values are not influenced by the prediction of the scale and depends only on the clustering and shape prototyping steps.

	RMSE		
	10 days	20 days	30 days
Case A (shape+scale)	2.81	3.85	4.97
Case B (scale)	22.99	29.59	7.51

The error rates computed during the ablation study are related to the inference of the shape prototypes and the scale values. The contribute of each inference on the measured performances depends on the definition of the prior Ground Truth knowledge of the specific experimental case under analysis, as detailed above. From the results of the above described experiments, we observed that the estimation of the popularity dynamics is more sensitive to errors related to the inference of the scale. In fact, an error on the scale estimation affects all the elements of the whole sequence and, hence, the RMSE value. Furthermore, since in our method we aim to predict the real values of the engagement without applying any quantization approach to the possible output values, there is not an upper bound for the estimated popularity scale. As consequence, the predicted scale values could be very large with high magnitude. Considering all the above, when the average error is computed, even a few large values could affect the final result. Indeed, by observing the distribution of the test errors, we found that the mean RMSE is skewed by a very few values that are larger than the others by several orders of magnitude. Figure 8 shows the box and whisker plot (i.e., boxplot) of the computed RMSE values for Case C (left plot in Figure 8). The distribution of the error values highlights that the mean of the RMSE errors is not a suitable metric for the evaluation of the results. Indeed, it is skewed by few large values.

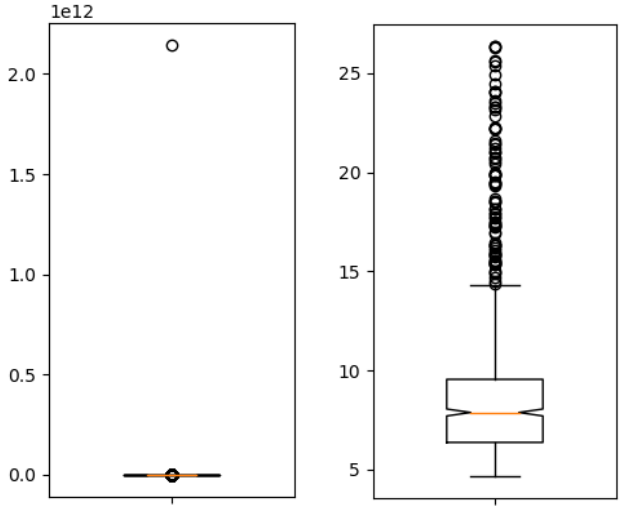

FIGURE 8. Box-and-whisker plot of an example of RMSE values computed on a test set (left). The mean RMSE value is skewed by only one value that is very large compared to others. The right plot shows the same distribution after removing the values lower than the first quartile and higher than the third quartile. In this example, the trimmed mean is 9, 00 and the median is 7, 87.

Furthermore, by analysing the distributions in Figure 4 is possible to observe the presence of several scale values (s_{scale}) that are outliers with respect their distribution (depicted as circles in the box-and-whisker plots). Indeed, the dataset includes few items with very large scale values. For example, there are only 11 photos which s_{scale} values are between 10.000 and 30.000 views, whereas the median values (the orange lines in Figure 4) are all lower than 100. As consequence, the presence of such few uncommon examples with very large values causes the skewness in the distribution of the test errors shown in Figure 8. For the above motivations, for the evaluation of the proposed method and the experiments of Case C (i.e., the only two settings in which the scale of the sequence is inferred) we considered two performance measures that are robust with respect to outlier values. In particular, we considered the 25% truncated RMSE (tRMSE 0.25) and the Median RMSE (RMSE MED). The truncated RMSE, also known as the interquartile mean, discards the same percentage of either high and low tails of a distribution. This means that both the best and the worst 25% performance values are ignored when the mean error is computed.

The right plot in Figure 8 shows the error distribution after removing the higher and lower 25% of values. After this process, the distribution of the errors is more clear, although it is still skewed by some outlier elements depicted as circles.

Algorithm 1 Popularity Sequence Prediction

Data: Input feature X, Ground Truth sequence s, shape prototype s^*_{shape}
Result: Inferred sequence $\hat{s}$
$n \leftarrow |s|$;
$\hat{c} \leftarrow RNDF.predict(X)$; // predict the cluster
$\hat{s}_{shape} \leftarrow getPrototype(\hat{c})$; // get the prototype sequence
$\hat{p} \leftarrow SVR.predict(X)$; // predict the popularity score
$\hat{s}_{scale} \leftarrow (e^{\hat{p}} - 1) \times n$;
$\hat{s} \leftarrow \hat{s}_{scale} \times \hat{s}_{shape}$; // predicted sequence

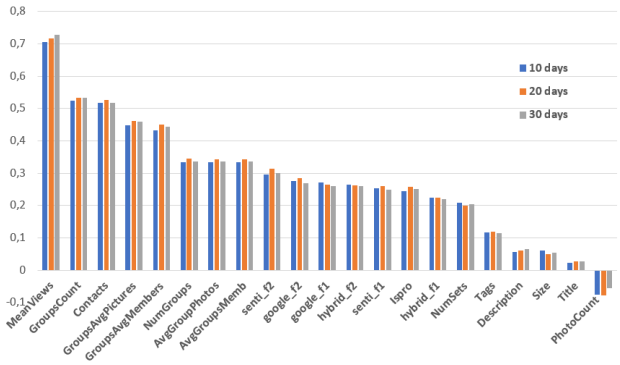

FIGURE 9. Spearman's correlation values between the $\hat{s}_{scale}$ estimation performed by a SVR trained with a single feature and the Ground Truth value, for the 10 days, 20 days and 30 days scenarios.

VI. POPULARITY DYNAMIC PREDICTION

Given a set of social features related to a Flickr photo, the proposed system uses a Random Forest classifier to predict the shape prototype of the sequence $\hat{s}_{shape}$. Then a SVR is employed to infer the popularity score [4] of the photo after n days. Then, Equation (4) is used to transform the output ot the SVR, to obtain the sequence scale estimation $\hat{s}_{scale}$. The final sequence $\hat{s}$ is obtained by combining the predicted shape $\hat{s}_{shape}$ and the inferred scale $\hat{s}_{scale}$ by applying the Equation (3) (Figure 3). The aforementioned procedure is described in the Algorithm 1. We repeated the whole pipeline (i.e., shape clustering, shape prediction, scale estimation) and the evaluation procedure considering $n = 10$, $n = 20$ and $n = 30$.

In the following paragraph, the performances of the proposed method on the prediction of the 30 days sequences are detailed. Then, the prediction of sequences with length 20 and 10 are discussed. Finally all the results are compared and commented.

A. PERFORMANCES ON 30 DAYS PREDICTION

The experimental results of the proposed method and of Case C are represented in Tables 2 and 3. As previously explained, the results are measured in terms of tRMSE and Median RMSE, at varying of the input features used by the SVR to perform scale regression. Table 2 shows the results obtained by training the SVR with a single feature. In particular, the fourth column in Table 2 details the Spearman's correlation values between the employed input feature and the popularity score.

As we can observe, the feature with the higher correlation is the mean number of user's photo views (MeanViews). The second ranked feature is the number of the groups the user is enrolled in (GroupsCount). Considering the achieved Spearman's values, is possible to notice that the features with higher correlation are the ones related to the users' statistics. The remaining columns in Table 2 report the trimmed RMSE (tRMSE) and Median RMSE (RMSE MED) error rates on the estimation of the whole sequence for the proposed method (columns 5 and 6) and the Case C (columns 7 and 8). Each row reports the results obtained by varying the features used in the estimation of the scale. The experimental results show that the best features are the MeanViews and GroupsCount, which allow to obtain evaluation performances very close to the Case C. Some interesting results have been obtained considering photo's features such as the number of groups the photo has been shared (NumGroups), and the statistics of such groups (AvgGroupsMemb and AvgGroupPhotos).

The experiments related to the single feature evaluation pointed out that the employed visual features achieves high error rates. In fact, the popularity of a photo in terms of number of views is directly related to the capability of the user and the photo to reach as many users as possible in the social platform (i.e., the user and photo potential audience in the social platform). Considering the results obtained in the evaluation of the single features (Table 2), we further considered specific compositions obtained by considering the features which obtained the best results for the estimation of the sequence scale in the single feature evaluation. To this aim, we evaluated several early and late fusion strategies. The early fusion consists on creating a new input for the SVR, obtained by the concatenation of the selected features. In particular, we evaluated 10 different combinations of features (see the first 10 rows in Table 3). For the sake of readability, each feature combination is assigned to an identifier (first column in Table 3). The achieved results show that the experiments which exploits features related to users' statistics obtains the lowest error rates. In fact, the higher error rates have been obtained by the combinations that do not include user features (i.e., combinations with IDs "photo" and "best_photo"). The best results in terms of tRMSE are obtained by combining the three best user's features (i.e., MeanViews, GroupsCount and Contacts) and the best photo's features (i.e., NumGroups, AvgGroupsMemb and AvgGroupsPhotos). Whereas the best results in terms of Median RMSE are obtained by using only the user related features (i.e., the concatenation of the features with indices from 0 to 6, identified by the ID "user" in Table 3).

We further evaluated the daily error rates obtained by the proposed approach. Instead of computing the mean error

TABLE 2. Results obtained by considering a single feature approach for the predictions of 30 days sequences. For each experimental setting, the fourth column reports the Spearman's correlation of the predicted s_{scale} value, the others columns report the trimmed RMSE (tRMSE 0.25) and the median RMSE (RMSE MED) considering the prediction of the whole temporal sequence. The results achieved by the proposed method are compared with respect to the Case C in which the shape prototype is known.

				Proposed Method		Case C (shape)	
Feature Source	Feature ID	Features	Spearman	tRMSE 0.25	RMSE MED	tRMSE 0.25	RMSE MED
User	0	Ispro	0,25	11,74	8,87	11,50	8,79
User	1	Contacts	0,52	10,48	8,96	10,22	8,69
User	2	PhotoCount	-0,06	11,00	9,65	10,70	9,31
User	3	MeanViews	**0,73**	**9,68**	**8,19**	9,44	7,95
User	4	GroupsCount	0,53	10,42	8,55	10,18	8,42
User	5	GroupsAvgMembers	0,44	11,19	9,37	10,91	9,11
User	6	GroupsAvgPictures	0,46	11,57	8,96	11,31	8,90
Photo	7	Size	0,05	11,03	9,45	10,72	9,18
Photo	8	Title	0,03	11,01	9,63	10,72	9,30
Photo	9	Description	0,06	10,96	9,58	10,66	9,21
Photo	10	NumSets	0,20	11,57	9,33	11,26	9,14
Photo	11	NumGroups	0,34	10,95	9,63	10,65	9,32
Photo	12	AvgGroupsMemb	0,34	10,88	9,54	10,59	9,24
Photo	13	AvgGroupPhotos	0,34	10,91	9,58	10,61	9,28
Photo	14	Tags	0,11	11,27	9,36	10,97	9,20
Visual	hybrid_f1	Hybridnet fc7	0,22	21,61	17,97	20,08	17,44
Visual	hybrid_f2	Hybridnet fc8a	0,26	13,37	11,76	12,95	11,38
Visual	senti_f1	DeepSentiBank fc7	0,25	21,16	18,27	20,60	17,82
Visual	senti_f2	DeepSentiBank fc8	0,30	16,52	14,32	15,99	13,77
Visual	google_f1	GoogleNet pool5/7x7_s1	0,26	13,85	12,15	13.43	11,74
Visual	google_f2	GoogleNet loss3/classifier	0,27	13,18	11,61	12,76	11,17

TABLE 3. Evaluation results for the prediction of the 30 days sequences obtained by combining the features.

			Proposed Method		Case C (shape)	
Feature ID	Features	Spearman	tRMSE 0.25	RMSE MED	tRMSE 0.25	RMSE MED
all	concat(0-14)	0,63	9,59	7,43	9,38	7,26
user	concat(0-6)	0,66	9,86	**7,03**	9,67	6,92
photo	concat(7-14)	0,28	10,80	8,58	10,50	8,30
best_photo	concat(11, 12, 13)	0,34	10,63	9,09	10,35	8,84
user2	concat(3,4)	0,71	9,60	7,66	9,38	7,54
user3	concat(1,3,4)	0,71	9,48	7,51	9,28	7,37
user5	concat(1, 3, 4, 5, 6)	0,68	9,77	7,30	9,59	7,24
	concat(user2, best_photo)	**0,72**	9,49	7,55	9,27	7,43
	concat(user3, best_photo)	0,71	**9,37**	7,38	9,16	7,27
	concat(user5,best_photo)	0,69	9,48	7,28	9,28	7,20
	Late Fusion 1 (AVG)	0,59	10,70	9,34	10,43	9,06
	Late Fusion 2 (SVR)	0,46	11,91	8,91	11,66	8,86
	Borda Count	0,63	10,60	9,16	10,33	8,87
	Weighted Fusion 2	0,71	9,88	8,34	9,63	8,18
	Weighted Fusion 3	0,71	9,95	8,49	9,69	8,26
	Weighted Fusion 4	0,68	10,30	8,65	10,05	8,49
	Weighted Fusion 5	0,67	10,44	8,79	10,18	8,63
	Weighted Fusion 6	0,67	10,47	8,88	10,20	8,67
	Weighted Fusion 7	0,67	10,51	8,92	10,24	8,64
	Weighted Fusion 8	0,67	10,52	8,96	10,25	8,69
	Weighted Fusion 9	0,66	10,51	8,99	10,24	8,73
	Weighted Fusion 10	0,65	10,52	9,03	10,25	8,80
	Weighted Fusion 11	0,65	10,53	9,06	10,26	8,80
	Weighted Fusion 12	0,65	10,53	9,07	10,26	8,81
	Weighted Fusion 13	0,65	10,53	9,08	10,27	8,82
	Weighted Fusion 14	0,65	10,54	9,08	10,27	8,82
	Weighted Fusion 15	0,65	10,54	9,08	10,27	8,81

between the predicted sequence and the Ground Truth, the daily squared errors are collected, and the daily tRMSE is then computed for all the error rates of the same day. The feature ''MeanViews'' obtains the lower error over all the period of observation by a certain margin. The features ''GroupsCount'' and ''Contacts'' achieves similar results in the early period (i.e., the first week), then their performances become slightly different. Substantially, the daily evaluation confirms the results of in Table 2 and Table 3.

B. PERFORMANCES ON 10 AND 20 DAYS PREDICTION

The resulting clusters in Figure 5 shows that the most of the sequences have a flat shape in the last days. By the other end, we can observe an higher shape variability in the first days. Such observation has been confirmed in the cluster analysis (see Section IV), which suggest a slightly lower number of clusters for the sequences with length 10 and 20 (45 clusters) with respect to the sequences with length 30 (50 clusters).

TABLE 4. Evaluation results for the prediction of the 10 days sequences.

				Proposed Method		Case C (shape)	
Feature Source	Feature ID	Features	Spearman	tRMSE 0.25	RMSE MED	tRMSE 0.25	RMSE MED
User	0	Ispro	0,24	9,12	**6,66**	8,18	6,29
User	1	Contacts	0,52	7,88	7,08	7,33	6,41
User	2	PhotoCount	-0,08	8,33	7,65	7,72	6,78
User	3	MeanViews	**0,71**	**7,60**	6,84	7,02	6,08
User	4	GroupsCount	0,53	7,93	6,89	7,36	6,12
User	5	GroupsAvgMembers	0,43	8,43	7,34	7,80	6,62
User	6	GroupsAvgPictures	0,45	8,77	7,24	8,11	6,43
Photo	7	Size	0,06	8,38	7,53	7,72	6,69
Photo	8	Title	0,02	8,31	7,64	7,72	6,79
Photo	9	Description	0,06	8,30	7,63	7,71	6,76
Photo	10	NumSets	0,21	8,72	6,68	7,90	6,49
Photo	11	NumGroups	0,33	8,33	7,65	7,72	6,78
Photo	12	AvgGroupsMemb	0,33	8,33	7,65	7,71	6,78
Photo	13	AvgGroupPhotos	0,33	8,34	7,65	7,72	6,78
Photo	14	Tags	0,12	8,54	7,54	7,88	6,66
Visual	hybrid_f1	Hybridnet fc7	0,22	16,01	13,96	13,75	11,68
Visual	hybrid_f2	Hybridnet fc8a	0,27	10,84	9,53	9,35	8,11
Visual	senti_f1	DeepSentiBank fc7	0,25	16,74	14,12	14,24	11,91
Visual	senti_f2	DeepSentiBank fc8	0,30	13,45	11,57	11,39	9,70
Visual	google_f1	GoogleNet pool5/7x7_s1	0,27	11,38	10,19	9,83	8,66
Visual	google_f2	GoogleNet loss3/classifier	0,28	10,71	9,58	9,33	8,14

TABLE 5. Evaluation results for the prediction of the 10 days sequences obtained by combining the features.

			Proposed Method		Case C (shape)	
Feature ID	Features	Spearman	tRMSE 0.25	RMSE MED	tRMSE 0.25	RMSE MED
all	concat(0-14)	0,61	7,73	6,18	6,84	5,37
user	concat(0-6)	0,64	8,00	6,00	7,11	5,48
photo	concat(7-14)	0,28	8,34	6,84	7,58	6,10
best_photo	concat(11,12,13)	0,33	8,10	7,30	7,54	6,59
user2	concat(3,4)	0,69	7,56	6,44	6,93	5,66
user3	concat(1,3,4)	0,69	7,50	6,37	6,84	5,63
user5	concat(1,3,4,5,6)	0,66	7,82	5,92	7,07	5,61
	concat(user2,best_photo)	**0,70**	7,46	6,41	6,86	5,60
	concat(user3,best_photo)	0,69	**7,39**	6,37	6,78	5,59
	concat(user5,best_photo)	0,67	7,53	**5,91**	6,86	5,55
	Late Fusion 1 (AVG)	0,57	8,20	7,53	7,53	6,68
	Late Fusion 2 (SVR)	0,47	9,45	6,87	8,37	6,62
	Borda Count (wAVG)	0,62	8,16	7,39	7,48	6,60
	Weighted Fusion 2	0,70	7,74	6,93	7,11	6,15
	Weighted Fusion 3	0,69	7,77	6,95	7,13	6,23
	Weighted Fusion 4	0,66	8,06	7,15	7,38	6,30
	Weighted Fusion 5	0,65	8,11	7,20	7,43	6,35
	Weighted Fusion 6	0,65	8,13	7,22	7,46	6,45
	Weighted Fusion 7	0,66	8,14	7,25	7,48	6,54
	Weighted Fusion 8	0,66	8,15	7,28	7,50	6,59
	Weighted Fusion 9	0,64	8,14	7,31	7,46	6,55
	Weighted Fusion 10	0,63	8,13	7,33	7,45	6,53
	Weighted Fusion 11	0,63	8,13	7,35	7,45	6,54
	Weighted Fusion 12	0,63	8,13	7,35	7,45	6,54
	Weighted Fusion 13	0,63	8,13	7,34	7,46	6,55
	Weighted Fusion 14	0,63	8,13	7,34	7,46	6,55
	Weighted Fusion 15	0,63	8,13	7,35	7,45	6,54

For this reason, we further evaluated our system on the sequences with length of 10 and 20 days. Note that both groups of data include the sequences of the 30 days set. Figure 9 shows the Spearman's correlations values between the $\hat{s}_{scale}$ estimated by the SVR properly trained with a specific feature and the Ground Truth s_{scale} value, for the 10 days, 20 days and 30 days scenarios. The features are ranked by the mean correlation among the three cases. As we can observe, the ranking of the features is similar for the three scenarios. Most of the user's features (MeanViews, GroupsCount, Contacts, GroupsAvgPictures and GroupsAvgMembers) achieve the highest correlations values. Among the photo's feature, the ones with highest performances are the features related to statistics of the groups in which the photo is shared in (i.e., NumGroups, AvgGroupPhotos and AvgGroupMemb). The PhotoCount feature (i.e., number of photos of the user) is the only one that achieves a negative value of correlation. The visual features are placed in the middle positions of the ranking.

TABLE 6. Evaluation results for the prediction of the 20 days sequences.

				Proposed Method		Case C (shape)	
Feature Source	Feature ID	Features	Spearman	tRMSE 0.25	RMSE MED	tRMSE 0.25	RMSE MED
User	0	Ispro	0,26	11,39	8,76	10,13	7,83
User	1	Contacts	0,53	9,73	8,67	8,96	7,70
User	2	PhotoCount	-0,08	10,25	9,68	9,43	8,31
User	3	MeanViews	**0,72**	**9,11**	**8,10**	8,32	7,16
User	4	GroupsCount	0,53	9,73	8,30	8,96	7,53
User	5	GroupsAvgMembers	0,45	10,54	9,03	9,61	8,22
User	6	GroupsAvgPictures	0,46	10,93	8,74	10,01	7,78
Photo	7	Size	0,05	10,26	9,37	9,39	8,27
Photo	8	Title	0,03	10,21	9,44	9,42	8,24
Photo	9	Description	0,06	10,19	9,50	9,39	8,22
Photo	10	NumSets	0,20	11,04	8,77	9,89	8,12
Photo	11	NumGroups	0,34	10,24	9,74	9,39	8,33
Photo	12	AvgGroupsMemb	0,34	9,91	9,07	9,18	7,91
Photo	13	AvgGroupPhotos	0,34	9,87	8,97	9,16	7,86
Photo	14	Tags	0,12	10,54	9,30	9,64	8,31
Visual	hybrid_f1	Hybridnet fc7	0,22	20,49	18,06	17,43	14,93
Visual	hybrid_f2	Hybridnet fc8a	0,26	13,86	12,28	11,79	10,27
Visual	senti_f1	DeepSentiBank fc7	0,26	21,45	18,67	18,03	15,49
Visual	senti_f2	DeepSentiBank fc8	0,31	16,45	14,28	13,91	11,74
Visual	google_f1	GoogleNet pool5/7x7_s1	0,26	14,07	12,63	11,98	10,37
Visual	google_f2	GoogleNet loss3/classifier	0,28	13,22	11,84	11,29	9,71

TABLE 7. Evaluation results for the prediction of the 20 days sequences obtained by combining the features.

			Proposed Method		Case C (shape)	
Feature ID	Features	Spearman	tRMSE 0.25	RMSE MED	tRMSE 0.25	RMSE MED
all	concat(0-14)	0,62	9,16	7,57	8,16	6,37
user	concat(0-6)	0,66	9,44	7,02	8,42	6,34
photo	concat(7-14)	0,28	10,10	8,39	9,12	7,38
best_photo	concat(11,12,13)	0,33	9,74	8,78	9,08	7,87
user2	concat(3,4)	**0,71**	8,96	7,63	8,22	6,78
user3	concat(1,3,4)	**0,71**	8,88	7,52	8,15	6,67
user5	concat(1,3,4,5,6)	0,68	9,23	**7,01**	8,41	6,49
	concat(user2,best_photo)	**0,71**	8,87	7,56	8,15	6,75
	concat(user3,best_photo)	**0,71**	**8,80**	7,48	8,10	6,66
	concat(user5,best_photo)	0,68	8,98	6,99	8,21	6,42
	Late Fusion 1 (AVG)	0,59	10,03	9,14	9,16	8,09
	Late Fusion 2 (SVR)	0,47	11,34	8,11	10,11	7,71
	Borda Count (wAVG)	0,63	9,96	8,94	9,09	7,95
	Weighted Fusion 2	**0,71**	9,27	8,24	8,52	7,34
	Weighted Fusion 3	0,70	9,39	8,35	8,60	7,48
	Weighted Fusion 4	0,67	9,75	8,57	8,92	7,64
	Weighted Fusion 5	0,67	9,88	8,67	9,03	7,71
	Weighted Fusion 6	0,67	9,90	8,70	9,06	7,74
	Weighted Fusion 7	0,67	9,90	8,71	9,06	7,77
	Weighted Fusion 8	0,67	9,91	8,75	9,06	7,75
	Weighted Fusion 9	0,66	9,90	8,79	9,05	7,83
	Weighted Fusion 10	0,65	9,90	8,84	9,05	7,87
	Weighted Fusion 11	0,65	9,91	8,85	9,06	7,88
	Weighted Fusion 12	0,65	9,91	8,86	9,06	7,89
	Weighted Fusion 13	0,65	9,91	8,86	9,06	7,89
	Weighted Fusion 14	0,65	9,92	8,87	9,06	7,89
	Weighted Fusion 15	0,65	9,91	8,87	9,06	7,89

Table 6 and Table 4 show the experimental results obtained by training an SVR on single features, on dataset of 20 days and 10 days sequences respectively. The experimental results in terms of Trimmed RMSE (tRMSE) and Median RMSE (RMSE MED) confirm the behaviour previously observed in the prediction of the 30 days sequences, and suggested by the correlation analysis. Indeed, user's features such as MeanViews, Contacts and GroupsCount resulted useful to achieve good performances. Among the photo's features, the NumSets achieves good results in terms of RMSE MED, whereas visual features don't obtain good performances.

Table 7 and Table 5 show the experimental results obtained by combining the input features, on the 20 days and 10 days datasets respectively. Also in these cases, proper combinations of features help to further improve the obtained results. In either the 10 days and the 20 scenarios, the best results have been obtained by combining the best user's features (i.e., Contacts, MeanViews, GroupsCount, GrpipsAvgMembers and GroupsAvgPictures) and the best

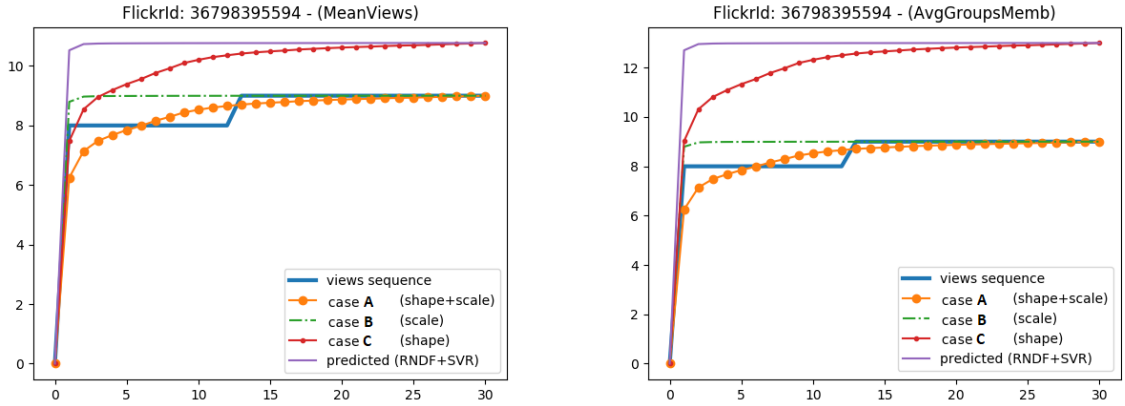

FIGURE 10. Inference of the same sequence obtained by using two single feature approaches. The left figure shows the result obtained by exploiting the "MeanViews" feature (i.e, the best feature according to Table 2), whereas the right figure shows the result obtained by exploiting the feature "AvgGroupsMemb" (i.e., the best single feature among the ones extracted from the photo information). The plots compare the sequences produced by the cases defined by the ablation study and by the proposed method with respect to the Ground Truth sequence. Best seen in color.

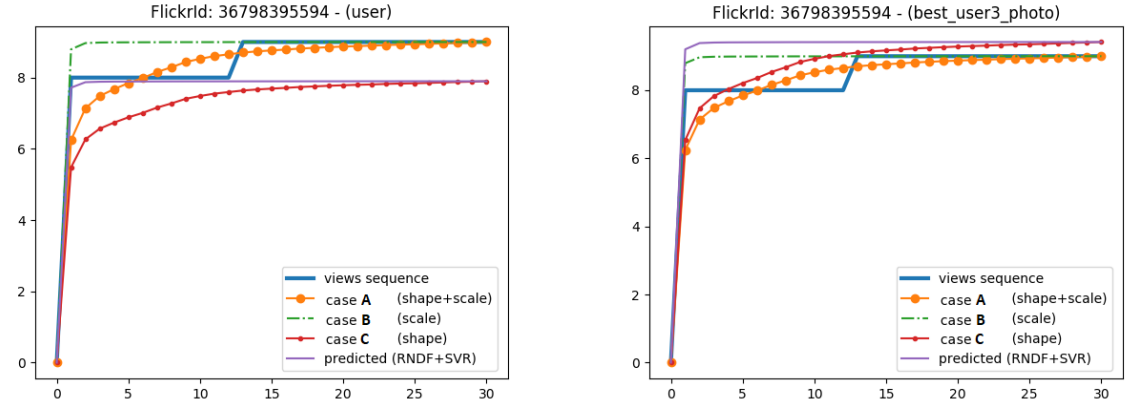

FIGURE 11. Inference of the same sequence obtained by using two combined feature approaches. The left figure shows the result obtained by exploiting the "user" feature combination, which is the best approach with respect to the "RMSE MED" measure (see Table 3). The right figure shows the result obtained by exploiting the combination of the best three user features (i.e., Contacts, MeanViews, and GroupsCount) and the best photo features (i.e., NumGroups, AvgGroupsMemb, AvgGroupPhotos), which is the best approach with respect to the "RMSE 0.25" measure (see Table 3). The plots compare the sequences produced by the cases defined in the ablation study and by the proposed method with respect to the Ground Truth sequence. Best seen in color.

photo features (i.e., NumGroups, AvgGroupsMemb and AvgGroupPhotos).

VII. CONCLUSION AND FUTURE WORK

The work presented in this paper investigates the challenging task of predicting the evolution of popularity (i.e. popularity dynamics) of photos shared through a social media platform. To benchmark the problem, a recently released public dataset is employed [21]. The paper also describes the experimental evaluation of a method that aims to predict the sequence of views over a period of 30 days of by a photo shared on Flickr, without constrains on the estimated values (e.g., coarse-quantization, upper bound threshold, etc.), nor considering the early values of the sequences. In particular, the proposed approach combines the results obtained by two different algorithms aimed to estimate the maximum number of the number of views reached by the photo in the period of observation (scale) and the shape of the sequence. The independent relationship between the two properties has been demonstrated. Furthermore, an ablation analysis of the problem aimed to understand how the single predictions affect the whole pipeline has been performed, this allowed the choice of proper metrics for the results evaluation, that is a crucial aspect in Machine Learning pipeline design. The proposed approach has been evaluated for the prediction of 10 days and 20 days sequences, which are characterized by an higher variability with respect to the 30 days sequences. The obtained results show that the proposed method obtains performance very close to Case C, in which the Ground Truth shape of the sequence is known. Figure 10 and Figure 11 compare the sequences obtained by the proposed method and by the cases defined during the ablation study with the Ground Truth

sequence, for the same engagement sequence example. In particular, these figures allow the comparison of the results obtained by the best single feature and the best combined feature approaches respectively.

We empirically observed that the considered visual features are not meaningful for the popularity prediction. However, previous works demonstrated that the visual appearance as well as the semantic content of a picture have an effect on the viewer emotional sphere [30]. Therefore, future works could investigate the visual aspect, trying to extract or learn proper features useful for this task. Future efforts will be devoted to the extension of the dataset by taking into account other social platforms (e.g., Twitter, Facebook), and if the specific preprocessing performed by the social network platform [35] has an influence in the achieved popularity. Furthermore, additional time-aware features can be considered, such as the day of the week and the hour of the day. Also, different approaches to treat the problem of popularity dynamics prediction as a time series forecasting task can be taken into account, as well as the extension of the experiments to the prediction of the number of favorites and the number of comments included in the built dataset.

REFERENCES

[1] R. Hanna, A. Rohm, and V. L. Crittenden, "We're all connected: The power of the social media ecosystem," *Bus. Horizons*, vol. 54, no. 3, pp. 265–273, May/Jun. 2011.

[2] A. Goeldi, "Website network and advertisement analysis using analytic measurement of online social media content," U.S. Patent 7 974 983, Jul. 5, 2011.

[3] M. Trusov, R. E. Bucklin, and K. Pauwels, "Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking site," *J. Marketing*, vol. 73, no. 5, pp. 90–102, Sep. 2009.

[4] A. Khosla, A. Das Sarma, and R. Hamid, "What makes an image popular?" in *Proc. 23rd Int. Conf. World Wide Web*, Apr. 2014, pp. 867–876.

[5] A. Ortis, G. M. Farinella, V. D'amico, L. Addesso, G. Torrisi, and S. Battiato, "Recfusion: Automatic video curation driven by visual content popularity," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1179–1182.

[6] S. Battiato, G. M. Farinella, F. L. M. Milotta, A. Ortis, L. Addesso, A. Casella, V. D'Amico, and G. Torrisi, "The social picture," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2016, pp. 397–400.

[7] T. Yamasaki, S. Sano, and K. Aizawa, "Social popularity score: Predicting numbers of views, comments, and favorites of social photos using only annotations," in *Proc. 1st Int. Workshop Internet-Scale Multimedia Manage.*, Nov. 2014, pp. 3–8.

[8] S. Cappallo, T. Mensink, and C. G. M. Snoek, "Latent factors of visual popularity prediction," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 195–202.

[9] F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo, and S.-F. Chang, "Image popularity prediction in social media using sentiment and context features," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 907–910.

[10] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 223–232.

[11] S. Aloufi, S. Zhu, and A. El Saddik, "On the prediction of flickr image popularity by analyzing heterogeneous social sensory data," *Sensors*, vol. 17, no. 3, p. 631, Mar. 2017.

[12] J. Hu, T. Yamasaki, and K. Aizawa, "Multimodal learning for image popularity prediction on social media," in *Proc. IEEE Int. Conf. Consum. Electron.-Taiwan (ICCE-TW)*, May 2016, pp. 1–2.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[14] A. Zohourian, H. Sajedi, and A. Yavary, "Popularity prediction of images and videos on instagram," in *Proc. 4th Int. Conf. Web Res. (ICWR)*, Apr. 2018, pp. 111–117.

[15] K. Almgren, J. Lee, and M. Kim, "Prediction of image popularity over time on social media networks," in *Proc. Annu. Connecticut Conf. Ind. Electron., Technol. Automat. (CT-IETA)*, Oct. 2016, pp. 1–6.

[16] K. Almgren, J. Lee, and M. Kim, "Predicting the future popularity of images on social networks," in *Proc. 3rd Multidisciplinary Int. Social Netw. Conf. Social Inform.*, Sep. 2016, Art. no. 15.

[17] L. Li, R. Situ, J. Gao, Z. Yang, and W. Liu, "A hybrid model combining convolutional neural network with xgboost for predicting social media popularity," in *Proc. ACM Multimedia Conf.*, Oct. 2017, pp. 1912–1917.

[18] B. Wu, W.-H. Cheng, Y. Zhang, and T. Mei, "Time matters: Multi-scale temporalization of social media popularity," in *Proc. ACM Multimedia Conf.*, Oct. 2016, pp. 1336–1344.

[19] B. Wu, T. Mei, W.-H. Cheng, and Y. Zhang, "Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition," in *Proc. AAAI*, Feb. 2016, pp. 272–278.

[20] B. Wu, W.-H. Cheng, Y. Zhang, Q. Huang, J. Li, and T. Mei, "Sequential prediction of social media popularity with deep temporal context networks," Dec. 2017, *arXiv:1712.04443*. [Online]. Available: https://arxiv.org/abs/1712.04443

[21] A. Ortis, G. M. Farinella, and S. Battiato, "Prediction of social image popularity dynamics," in *Image Analysis and Processing—ICIAP 2019* (Lecture Notes in Computer Science), vol. 11752, E. Ricci, S. R. Bulò, C. Snoek, O. Lanz, S. Messelodi, and N. Sebe, Eds. Cham, Switzerland: Springer, 2019.

[22] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini, "A peek into the future: Predicting the evolution of popularity in user generated content," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, Feb. 2013, pp. 607–616.

[23] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of youtube videos," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, Feb. 2013, pp. 365–374.

[24] G. Szabo and B. A. Huberman, "Predicting the popularity of Online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[25] S. D. Roy, T. Mei, W. Zeng, and S. Li, "Towards cross-domain learning for social video popularity prediction," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1255–1267, Oct. 2013.

[26] J. Wu, Y. Zhou, D. M. Chiu, and Z. Zhu, "Modeling dynamics of Online video popularity," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1882–1895, Sep. 2016.

[27] T. Trzciński and P. Rokita, "Predicting popularity of Online videos using support vector regression," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2561–2570, Nov. 2017.

[28] R. G. Garroppo, M. Ahmed, S. Niccolini, and M. Dusi, "A vocabulary for growth: Topic modeling of content popularity evolution," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2683–2692, Oct. 2018.

[29] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," in *Proc. ICWSM*, vol. 12, May 2012, pp. 26–33.

[30] A. Ortis, G. M. Farinella, and S. Battiato, "An overview on image sentiment analysis: Methods, datasets and current challenges," in *Proc. 16th Int. Joint Conf. E-Bus. Telecommun.*, vol. 1, 2019, pp. 290–300.

[31] M. Valafar, R. Rejaie, and W. Willinger, "Beyond friendship graphs: A study of user interactions in Flickr," in *Proc. 2nd ACM Workshop Online Social Netw.*, Aug. 2009, pp. 25–30.

[32] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[33] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," Oct. 2014, *arXiv:1410.8586*. [Online]. Available: https://arxiv.org/abs/1410.8586

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[35] O. Giudice, A. Paratore, M. Moltisanti, and S. Battiato, "A classification engine for image ballistics of social data," in *Image Analysis and Processing*, S. Battiato, G. Gallo, R. Schettini, and F. Stanco, Eds. Cham, Switzerland: Springer, 2017, pp. 625–636.

ALESSANDRO ORTIS received the master's degree *(summa cum laude)* in computer science from the Universitá degli Studi di Catania, in 2015, and the Ph.D. degree in mathematics and computer science from TIM, in 2019. His Ph.D. thesis investigates several aspects related to visual sentiment analysis applied on crowd sourced images/videos. He has been involving in the field of computer vision research, since 2012, when he joined the Image Processing Laboratory (IPLab). He did two research internships with STMicroelectronics, in 2011/2012 and with TIM, in 2015. He is currently a Postdoctoral Researcher with the Universitá degli Studi di Catania. He is the coauthor of 15 articles published in international conferences and four journals and co-inventor of one international patent. His current research interests include computer vision, machine learning, and multimedia. He received the Archimede Prize for the excellence of academic career conferred by the University of Catania, in 2015. He is a reviewer of several international conferences and journals.

GIOVANNI MARIA FARINELLA (M'11–SM'16) is currently a tenure track Associate Professor with the Department of Mathematics and Computer Science, Universitá degli Studi di Catania, Italy. He is the author of more than 100 articles in international book chapters, journals, and conference proceedings, and the co-inventor of five patents involving industrial partners. His current research interests include computer vision, pattern recognition, and machine learning. His group's most recent effort is related to first person egocentric vision. Dr. Farinella serves as a reviewer and on the board programme committee for major international journals and conferences, including CVPR, ICCV, ECCV, and BMVC. He received the PAMI Mark Everingham Prize, in 2017. He is an Area Chair of ICCV 2017–2019. He founded the International Computer Vision Summer School (ICVSS), in 2006, which he currently directs. He also founded the Medical Imaging Summer School (MISS), in 2014, which he currently directs. He has been an Associate Editor of the international journal *Pattern Recognition*, since 2017.

SEBASTIANO BATTIATO (M'04–SM'06) received the degree *(summa cum laude)* in computer science from the Universitá degli Studi di Catania, in 1995, and the Ph.D. degree in computer science and applied mathematics from the University of Naples, in 1999. From 1999 to 2003, he was the Leader of the "Imaging" Team, STMicroelectronics, Catania. He joined the Department of Mathematics and Computer Science, Universitá degli Studi di Catania, as an Assistant Professor, an Associate Professor, and a Full Professor, in 2004, 2011, and 2016, respectively. He has been the Chairman of the Undergraduate Program in Computer Science, from 2012 to 2017, and a Rector's Delegate of education (Postgraduates and Ph.D.), from 2013 to 2016. He is currently a Full Professor of computer science with the Universitá degli Studi di Catania, where he is also the Scientific Coordinator of the Ph.D. Program in Computer Science. He is involved in the research and directorship with the Image Processing Laboratory (IPLab). He coordinates IPLab's participates on large scale projects funded by national and international funding bodies and private companies. He has edited six books and coauthored about 200 articles in international journals, conference proceedings, and book chapters. He is a co-inventor of 22 international patents. His current research interests include computer vision, imaging technology, and multimedia forensics. Prof. Battiato has been a regular member of numerous international conference committees. He was a recipient of the 2017 PAMI Mark Everingham Prize for the series of annual ICVSS schools and the 2011 Best Associate Editor Award of the IEEE Transactions on Circuits and Systems for Video Technology. He has been the Chair of several international events, including ICIAP 2017, VINEPA 2016, ACIVS 2015, VAAM2014-2015-2016, VISAPP2012-2015, IWCV2012, ECCV2012, ICIAP 2011, ACM MiFor 2010-2011, and SPIE EI Digital Photography 2011-2012-2013. He has been a guest editor of several special issues published in international journals. He is an Associate Editor of the *SPIE Journal of Electronic Imaging* and the *IET Image Processing* journal. He is the Director (and Co-Founder) of the International Computer Vision Summer School (ICVSS). He is a reviewer of several international journals. He participated as a principal investigator in many international and national research projects.